The multiple life cycles of open data creation and use

Charalabidis, Yannis; Zuiderwijk, Anneke; Alexopoulos, Charalampos; Janssen, Marijn; Lampoltshammer, Thomas; Ferro, Enrico

DOI
10.1007/978-3-319-90850-2_2

Publication date
2018

Document Version
Final published version

Published in
Public Administration and Information Technology

Citation (APA)
Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). The multiple life cycles of open data creation and use. In *Public Administration and Information Technology* (pp. 11-31). (Public Administration and Information Technology; Vol. 28). Springer. https://doi.org/10.1007/978-3-319-90850-2_2

Chapter 2
The Multiple Life Cycles of Open Data Creation and Use

Open data can be defined as data that is free of charge or provided at marginal cost, under an open licence, machine readable, and provided in an open format

2.1 Introduction

Different terminologies have been suggested towards the description of various models of open data. The open data life cycle, the open data value chain or the open data process (Zuiderwijk, Janssen, Choenni, Meijer, & Alibaks, 2012) are terminologies illustrating different purposes – practical guidance or analytical understanding – and foci. Whereas value chain models – that will be further analysed in Chap. 7 – focus more on the creation of value during open data usage, the life cycle models aim to structure the handling of the data itself. Existing process models focus on activities within public administrations, such as generating (create/gather), editing (pre-process and curate) and publishing the data without paying too much attention on the outside-use and re-use processes.

In order to fully exploit the benefits of open data, traditional "one-way street" open data practices and initiatives should be replaced by an open data ecosystem, i.e. an approach to open data that focuses not only on data accessibility, but also on the larger environment for open data use—its "ecosystem" (Pollock, 2011; World Bank Group, 2015). An open data ecosystem can be defined as a cyclical, sustainable, demand-driven and environment-oriented around agents that are mutually interdependent in the creation and delivery of value from open data (Boley & Chang, 2007; Harrison, Pardo, & Cook, 2012; Heimstädt, Saunderson, & Heath, 2014).

Because of these many interdependencies, open data ecosystems should be studied as a whole, by investigating both the user and the publisher sides of the life cycle as well as the relation to each other. (Susha, Janssen, & Verhulst, 2017) in their proposal for a user-centric and interdisciplinary research agenda to advance open data: *"To realize its potential there is a need for more evidence on the full life cycle*

Y. Charalabidis et al., *The World of Open Data*, Public Administration and Information Technology 28, https://doi.org/10.1007/978-3-319-90850-2_2

of open data – within and across settings and sectors". In other terms, interdisciplinary open data research should investigate the open data life cycle in all its phases and address open data developments in different domains.

The open data life cycle is a conceptualization of the process and practices around handling data, starting from its creation, through the provision of open data to its use by various parties. In addition, the characteristics and interests of different stakeholders involved are hardly recognized and taken into account. Analysing different data life cycle models from technological (data curation, big data and linked data) and stakeholders (publishers and users) perspectives, this chapter introduces an advanced open data life cycle model based on all the above identifying associated tools for each stage of the cycle, as well as, the transitions and interdependencies between different phases.

Moreover, the advent of Linked and Big Data as well as the collaboration capabilities of Web 2.0 paradigm reformed the landscape of open data since they introduced enhanced capabilities. These advanced capabilities, in their turn, introduced different concepts, solutions and complexity in the data re-use, storing, analysis, and publication processes.

This chapter introduces the new requirements for open data provision and usage in terms of different technologies (linked and big data) along with the accompanying impediments as well as an overview of the existing life cycle models for open data in Sect. 2.2. Section 2.3 presents an accumulative model derived from the conjunction of the two different stakeholder sides as well as the duality of the users' roles in an open data ecosystem. It also defines different tools and methods in each step of the open data life cycle concerning the requirements of different types of data. Section 2.4 familiarizes different uses of the open data life cycle presenting the open data life cycle from the perspectives of the two different stakeholders, namely, the open data producer and the open data user. It also describes the application of the open data life cycle model in the research domain supporting the development of a Scientific Data Infrastructure (SDI). Finally, Sect. 2.5 concludes the chapter referring to the principles underpinning the life cycle and the open data ecosystem.

2.2 New Requirements for Open Data Provision and Usage

2.2.1 Linked Data

The linked data paradigm puts an emphasis on the structure of the data using triples and description based on RDF (Resource Description Framework) vocabularies as well as in storing technologies (SPARQL) solving also the issues of uniqueness and metadata. Linked data is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. The concept builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than

using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried (Soylu, Mödritscher, & De Causmaecker, 2012).

When we are dealing with linked data and since it is a quite novel technology, there are some important impediments that should be taken into account (Auer et al., 2012). First of all, linked data uses RDF Data Management Systems (i.e. SPARQL) which are more challenging than the relational data management. Ways of limiting this performance gap include column-storage technology, dynamic query optimization and other. Secondly, creating and maintaining links in a (semi-) automated fashion is still a major challenge and crucial for establishing coherence and facilitating data integration. New linking approaches should yield high precision and recall, which configure themselves automatically or with end-user feedback. Thirdly, since linked Data on the Web is mainly raw instance data, data integration, fusion, search and many other capabilities need to be linked and integrated with upper level ontologies. Fourthly, the quality of content on the Data Web varies, as the quality of content on the document web varies. Finally, since Data on the Web is dynamic, it is essential to facilitate the evolution of data while keeping things stable in methods development to spot problems in knowledge bases and to automatically suggest repair strategies. An example of linked data usage is presented in Sect. 2.4.4.

2.2.2 Big Data

The potential benefits of Big Data are significant, but many technical challenges should be addressed to fully accomplish those benefits (Jagadish et al., 2014). One of the most renowned challenges is the sheer size of the data. However, there are others such as Variety and Velocity completing the 3 V's of big data. Variety refers to heterogeneity of data types (structured and unstructured) originated by disperse data sources aiming at data representation and semantic interpretation. Velocity implies the time frame the data should be analyzed according to the rate of data arrival. Further important requirements have been detected since big data applications began such as veracity (reliability), variability (complexity) (Gandomi & Haider, 2015), privacy and usability (Jagadish et al., 2014).

Dealing with big data is a quite exhaustive task bringing in changes in technological and analytical level of data processing as well as in data storage with the most prominent technology to be the NoSQL databases. The advent of big data alternates the importance of the life cycle steps placing more focus on the "create", "process" and "store" steps of the life cycle. Technologies for covering these steps are the major concern at the moment. New analysis methods (indexing algorithms towards timely data analysis) have derived and applied on big data. An example of big data usage is presented in Sect. 2.4.2.

2.2.3 Web 2.0

In addition following the Web 2.0 paradigm (Alexopoulos, Loukis, & Charalabidis, 2014; Charalabidis, Alexopoulos, & Loukis, 2016) there is a new generation of OGD platforms and virtual environments trying to fill the gap of communication between data users and data providers through closing the feedback loop as well as creating the notion of data 'pro-sumers'. This shifts the paradigm towards highly active users, who assess the quality of the data they consume and mention weaknesses of them and new needs they have; who often become both consumers and providers of data is characterised by advanced capabilities to data users for commenting, rating, processing in order to improve them, adapt them to their specialized needs, or link them to other datasets (public or private); and then uploading-publishing new versions of them, or even their own new datasets. This systemic view of open data could be used to the development of new solutions matching supply and demand and utilising the innovation aspect of open data.

Zuiderwijk, Loukis, Alexopoulos, Janssen, and Jeffery (2014) proposed an open data electronic marketplace with enhanced capabilities for both producers and users. The new marketplace also supports the data pro-cumer enabling advanced publication procedures connected with the appropriate tools. The EU-FP7-ENGAGE project could be seen as such a marketplace, since its functionality supports all the identified requirements except the payment and value definition procedures which have not been realised in the ENGAGE context. Without the value definition and payment procedures the ENGAGE platform could be seen as a crowdsourcing-based platform for data processing and data exchange among users. The basic and novel functionality of such an architecture is shown in Table 2.1.

Table 2.1 Classical and novel functionality of OGD infrastructures adapted by Zuiderwijk et al. (2014)

Functionality	Stakeholder	Description
Classical open data functionality		
Data Publication	Provider	Support for publication to the providers: tutorials and guiding principles for data uploading
Data Modeling	Provider	Capabilities of flat metadata descriptions (based on a specific metadata models) and data formats
Data Search	User	Simple search via keywords, resource format, publisher, topic categories and countries
Data Visualisation	User	Simple visualisation techniques on specific datasets (maps, charts)
Data Download	User	Data and metadata downloading capabilities. Provision of API.

(continued)

Table 2.1 (continued)

Functionality	Stakeholder	Description
Novel open data functionality		
Grouping and Interaction	Provider/ User	Capabilities for (a) searching for and finding other users/ providers having similar interests in order to have in-formation and knowledge exchange and cooperation, (b) forming groups with other users/providers having similar interests in order to have information and knowledge exchange and cooperation, (c) maintaining datasets/working on datasets within one group, (d) communicating with other users/providers through messages in order to exchange information and knowledge and (e) getting immediately updated about the upload of new versions and enrichments of datasets maintained/worked on within the group, or new relevant items (e.g. publications, visualizations, etc.).
Data Processing	Provider/ User	Capabilities for (a) data enrichment – i.e. adding new elements – fields, (b) for metadata enrichment – i.e. fill in missing fields, (c) for data cleansing – e.g. detecting and correcting ubiquities in a dataset, matching text names to database IDs (keys) etc., (d) converting datasets to another format, (e) submitting various types of items – e.g. visualisations, publications – related to a dataset and (f) datasets combination and Mash-ups.
Data Enhanced Modeling	Provider/ User	Capabilities for description of flat, con-textual and detailed metadata of any metadata/vocabulary model.
Feedback and Collaboration	Provider/ User	Capabilities (a) to communicate own thoughts and ideas on the datasets to the other users and the providers of them through comments, (b) to read interesting thoughts and ideas of other users on the datasets through comments they enter on them, (c) to express our own needs for additional datasets that would be interesting and useful to me, (d) to get informed about the needs of other users for additional datasets and (e) to get informed about datasets extensions and revisions.
Data Quality Rating	User	Rating system against the basic quality aspects of datasets with capabilities to: (a) get informed on the level of quality of the datasets perceived by other users through their ratings and (b) communicate to the other users and the providers the level of quality of the datasets that I perceive.
Data Linking	Provider/ User	Capabilities of data and metadata linking to other ontologies in the web of data (Linked Open Data Cloud). Capabilities of querying data and metadata through SPARQL endpoints.
Data Versions Publication	Provider/ User	Support for publication/upload of new versions of the existing datasets, and connection with previous ones and initial datasets.
Data Visualisation	User	Advanced visualization techniques and visual analytics on specific datasets and/or datasets mashups (maps, charts, plots, series and other)

2.2.4 Models Describing the Data Life Cycle

Most models contain similar elements and differ only regarding semantics, granularity or the extension of the process (Carrara, Fischer, & Steenbergen, 2015). As a first remark emerging from the analysis of Table 2.2, the existence of a perfect life cycle model is not possible based on the various aspects (i.e. curation, preservation) and unique characteristics in each type of data (i.e. linked, big). Different models could be more applicable in different contexts as it can be observed in the examples of Table 2.2.

It is also observed that there are a lot of common stages/steps/phases that could be considered neutral being present in most of the life-cycle models, such as: discovery and acquisition, data organization, publication, integration, analysis, re-use and storage/preservation. These models describe the life-cycle as a sequential, one-dimensional process of activities that an unspecified set of actors repeatedly undertake in order to provide a formerly unexposed amount of data to an abstract general public.

Whereas only making available large volumes of different types of data might result in searching for a needle in a hay stack, the use of predefined views and apps might filter too much information to deliver true transparency. Linked data could be referred as a technology that enables the connection of different datasets in the web of data, in which the searching, acquiring and analysis capabilities are more structured but not too effective. The connection is achieved through the modelling stage of the linked data life-cycle. The modelling stage utilizes vocabularies and generic ontologies (FOAF, SKOS, RDF) for the description of the data in order to establish linkages between different datasets.

Furthermore, these models include only one analytical level. They exclusively take the operational processes of open data publication into account (such as extracting, cleaning, publishing and maintaining data), while largely ignoring the strategic processes (such as policy production, decision-making and administrative enforcement). Thus, the decisions which data will be published, who extracts data, how are data edited, how data can be accessed, which licenses are available, how data privacy and liability issues are treated, and who is involved in these decisions remain underappreciated (Open Data Monitor, 2015).

The data curation model is the only model that could be considered as being comprehensive, since it includes administrative and managerial processes. These more general strategic processes about open data refer to the governance structure, likely to be connected to an organization's ICT and data governance. For example, the planning and the execution of preservation actions throughout the curation life-cycle of the digital material. This would include plans for management and administration of all curation activities in the life-cycle.

The outlined issues point to another blind spot of most open data life-cycle models that these are actor-blind. Until the final model for linked data (section) was conceptualized there were no feedback capabilities and limited capabilities on retrieving, integrate and re-use open data. If at all, institutional characteristics and

Table 2.2 Data life cycle models

Model	Key elements	Part of the open data life cycle covered	Strength(s) of this model	Weakness(es) of this model	Example of how this model can be used
DCC Curation Lifecycle Model	(a) the data itself divided in digital objects and databases, (b) administrative and managerial actions, (c) the basic model and (d) the evaluation actions	Create, Pre-process, Curate, Store, Acquire, Process, Use	Curation preservation of data + Managerial and Administrative Procedures	Ideal model, not very realistic	A generic data management model.
Villazon – Terrazas et al. (2011)	(1) Specify; (2) Model; (3) Generate; (4) Publish; (5) Exploit	Create, Curate, Publish, Use	Focused on linked data	Not applicable in other contexts. Very generic.	Could be used from linked data publishers supporting re-use. Only for managerial purposes.
Hyland et al. (2011)	(1) identify, (2) model, (3) name, (4) describe, (5) convert, (6) publish	Pre-process, Curate, Publish	Focused on linked data publication process	Not applicable in other contexts. No inclusion of managerial processes and definition of a data plan.	Could be used from linked data publishers.
Hausenblas and Karnstedt (2010)	adding the steps (7) discovery (8) integration (9) use cases	Acquire, Process, Use	Focused on linked data. Includes re-use and the user side	Not applicable in other contexts	Could be used from linked data publishers and users.
Open Data Support Working Group	(1) Select (2) Model (3) Publish (4) Find (5) Integrate (6) Re-use (7) feedback	Create, Curate, Publish, Use, Feedback	Feedback loop Matching supply and demand	Very abstract. No peculiarities are addressed.	Could be used from linked data publishers and users. Could be used from public administrations for managerial purposes.

(continued)

Table 2.2 (continued)

Model	Key elements	Part of the open data life cycle covered	Strength(s) of this model	Weakness(es) of this model	Example of how this model can be used
van den Broek et al. (2011)	(1) identification, (2) preparation, (3) publication, (4) re-use and (5) evaluation	Pre-process, Curate, Publish, Use, Half Feedback step	The evaluation procedure	Not very descriptive	Could be used from linked data publishers supporting re-use and evaluation. Only for managerial purposes.
Auer et al. (2012)	Manual Revision and Authoring; Interlinking and Fusing; Classification and Enrichment; Quality Analysis; Evolution and Repair; Search and Browsing; Extraction; Storing and Querying	Create, Pre-process, Curate, Process, Use	Very detailed description of linked data manipulation	No feedback and collaboration mechanisms.	Could be used from public administrations providing linked data as well as linked data users.
Erl, Khattak, and Buhler (2016)	Data Identification; Data Acquisition and Filtering; Data Extraction; Data Validation and Cleansing; Data Aggregation and Representation; Data Modelling and Analysis; Data Visualization	Acquire, Curate, Process, Use	Very detailed description of big data handling from the user side	No publication procedures. More focused in the business sector and internal data analysis	Could be used from big data analysts and big data scientists
Kucera (2015)	OD Initiative initiation; Goal Setting; Publication Plan; Preparation of Datasets and infrastructure; Publication; Archiving; Evaluation.	Publication	Focused on managerial processes of data publication including evaluation procedures	Most for OGD initiatives	Could be used from public administration for publishing their data through an Open data initiative.

(continued)

Table 2.2 (continued)

Model	Key elements	Part of the open data life cycle covered	Strength(s) of this model	Weakness(es) of this model	Example of how this model can be used
Demchenko, Grosso, De Laat, and Membrey (2013)	Experiment planning; Data Collection and filtering; Data analysis (scientific data production); Data Re-purpose; Publication of data; Archive (data and scientific paper);	Acquire, Process, Use, Store	Actor blind/ Pro-cumers	Focused on Scientific Data Lifecycle	Could be used from universities embracing the open data paradigm for their research data and information.

https://joinup.ec.europa.eu/sites/default/files/D2.1.1%20Training%20Module%202.1%20The%20Linked%20Open%20Government%20Data%20Lifecycle_v0.11_EN.pdf

actor-interests are considered as "impediments" (Zuiderwijk, Janssen, Choenni, et al., 2012) or restrictions hindering an inherently good and beneficial idea (Meijer, de Hoog, Van Twist, van der Steen, & Scherpenisse, 2014). This is especially relevant as the different stakeholders involved have different understandings of and interests in open data which in turn influences the results (Janssen & Zuiderwijk, 2014; Zuiderwijk & Janssen, 2014a). Efforts have thus been made to develop more holistic analytic perspectives on open data e.g. based on complexity theory (Meijer et al., 2014) and the information ecology approach (Harrison, Guerrero, et al., 2012).

2.3 The Open Data Life Cycle: An Ecosystem Approach

The ecosystem perspective is widely used by scholars, policy makers and other stakeholders across different domains to discuss and explore the interdependencies among data, technology, actors and innovation in several organizational and technological contexts (Harrison, Guerrero, et al., 2012). The added value of the ecosystem perspective on open data is its focus on the relationships and interdependencies between the social (publishers and users of open data) and technological (data linking, big data analysis, storing, visualising) factors that affect the performance of open data activities within the life cycle (Dawes, Vidiasova, & Parkhimovich, 2016).

Addressing the new requirements under the ecosystem concept, a hybrid model has been produced incorporating steps from all its predecessors (see Sect. 2.2.4). Various steps addressing linked and big data specific capabilities along

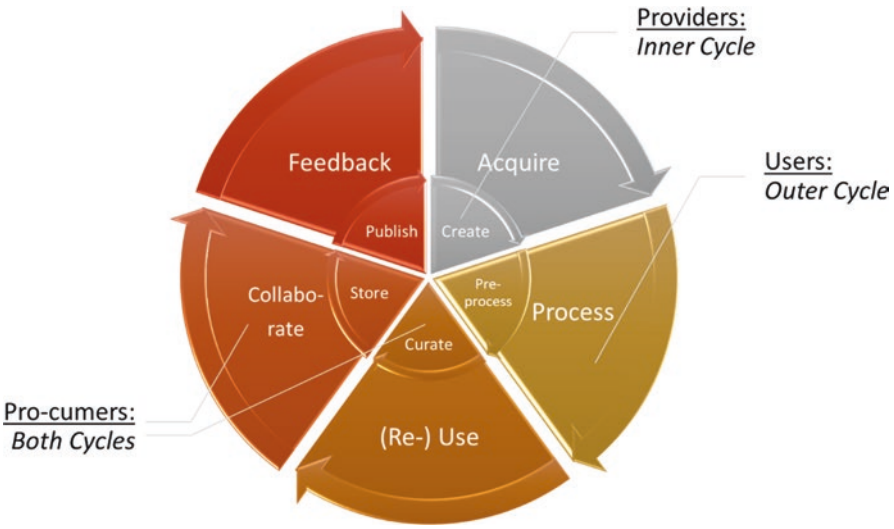

Fig. 2.1 The open data life cycle model

with the identification of the proper tools as well as the two different sides of the open data life cycle have been merged into a wider life cycle model providing the ecosystem view towards the achievement of the abovementioned impact from opening of public data. The curation life cycle is embedded in the "Curate" and "Pre-process" steps of the ENGAGE Open Data Life Cycle. Steps from the Open Data Publication Methodology (Kucera, 2015) have been also included. The basic development of the ENGAGE project since its conception is the collaboration step which is not included in any one of the above models. This is a result of the ENGAGE advanced functionality and web 2.0 capabilities which in fact provide a solid solution towards the realisation of the HORIZON 2020 vision concerning the e-infrastructures development for new workflows and collaboration.

Figure 2.1 introduces the Open Data Life Cycle Model. The different roles of the system are recognised in terms of inner and outer cycles. At this point we would like to clarify the pre-process step which is not referring to the calibration of data reducing their value. It incorporates the goal setting for each individual organisation publishing open data. The "Publish" step incorporates the publication planning which is related with the goals setting method of the "pre-processing" step. What is more, the feedback step refers to both the feedback from users as well as the assessment of the publication process against the goals setting.

Table 2.3 presents the methods and tools used for each life cycle stage regarding different types of data (big and linked).

Table 2.3 Methods and tools in each step of the open data life cycle

Life cycle stage	Tools	Methods
Create/Gather: The process of creating data	Sensors; RFID, IoT, IS; Human; Connection with already gathered open data; Hadoop for big data	Automated data creation (logs, network data) (Chen et al., 2014); Manual data entry; Linking with Open Data Portals
Pre-process: The managerial process of defining data quality	Detailed Metadata Standards; Evaluation Metrics and Models; Maturity Matrices; Unique identification (URIs and URLs)	Conceptualization & Goal setting; Evaluation plan and data quality; 3-layer Metadata Schema for portals
Curate: The process of meeting the required data quality and legal requirements	LOD Refine External Tool; Individual/Native Tools; R	Structuring; Anonymization; Metadata Refinement; Change Data Format; Data Cleansing
Store/Obtain: The decision making process of storing.	Data Centres; SPARQL Repositories for linked data; NoSQL & Document Databases for big data, linking with other datasets	Versioning; Data Linking; K-value and column oriented databases for big data (Chen et al., 2014)
Publish: The process covering legal issues	Upload Capability	Publication Plan Open Access Licensing Intellectual Property Rights
Retrieve/Acquire: The process of data acquisition through OD portals	OD portals (e.g. European data portal, world bank, national initiatives)	Multilingual search techniques APIs
Process: The process of data analysis	External data processing tools: Open Refine; R; Rapidminer; KNMINE; excel; Weka/ Pentaho	Data enrichment; Create Linked Open Data; Different Datasets combination; Text and Data Mining; Hashing; Cluster Analysis & Factor Analysis (Chen et al., 2014)
Use: The process of presenting the analysis outcomes	Internal & External Visualization tools; Statistical Packages; Linking with external artefacts (publications)	Statistical Analysis; Map Visualization; Chart Visualization; Plot Visualization; Visual Analytics; Cluster diagrams
Collaborate: The process of communicating with other data users	Collaboration space and workflow Web 2.0 capabilities and tools	Exchange notes/emails/ideas Create Groups of common interests
Feedback: The process of evaluating and providing feedback to data providers	Declare Need Web 2.0 Capabilities and Tools	Data Quality Rating; Requests on Open Data; Assessment of Publication

2.4 Different Uses of the Open Data Life Cycle

Much research has been conducted and many models have been designed in order to identify the open data life cycle as we can observe in Table 2.2. Each model focuses on different perspectives of open data regarding its nature (linked and big) and its purpose (data management, data curation). Even more research has been conducted for the definition of the data management life cycle (Committee on Earth Observation Satellites, Working Group on Information Systems and Services, 2011). This subsection analyses models that conceptualize the practices around handling data, from its generation to administrative practices involved in the provision of open data by public sector institutions to its use by third-parties.

This sub-section describes in more detail open data life cycle that best suits in different cases in order to illustrate specific aspects of the open data life cycle. As it could be discerned from the previous sub-sections the open data life cycle could be seen by two different perspectives. The major distinguishing aspect of the open data life cycle is the different stakeholders i.e. the publishers and the users. In the following sub-sections we present the open data life cycle from the publisher's side originating from the EU COSMODE project (Kucera, 2015) and the open data life cycle from the user's side. The user side consists of multiple stakeholders (i.e. scientists, journalists and citizens).

2.4.1 Towards Publication: The Data Publisher's Side

Open data are essential for achieving the United Nations' Sustainable Development Goals (The Open Working Group, 2015). Increased transparency, accountability and citizen participation (Jetzek, Avital, & Bjørn-Andersen, 2013), improved efficiency and effectiveness of public services (Huijboom, Broek, & Dutch Ministery of the Interior and Kingdom Relations, 2011), stimulation of economic growth; creation of social value (Gruen, Houghton, & Tooth, 2014) and positive impact on the quality and the effectiveness of the political debate (Ubaldi, 2013a), are only some examples of what our society could achieve through the opening and re-use of open data.

For the above-mentioned reasons, many countries all over the world design and implement OGD initiatives. Such initiatives have resulted in a greater availability of data including legislative interventions and development of digital infra-structures for this purpose (Commission of the European Communities, 2011). According to the Open Knowledge Network (2017), the "keep it simple" principle should be followed when opening up data. Even though OGD initiatives have been launched in many countries across the globe, only over 10% of the 1.290 datasets surveyed in the second edition of the Open Data Barometer study were published under an open license, in bulk and in machine-readable formats.

In addition, (Zuiderwijk, Janssen, Choenni, et al., 2012) observed that in practice it might be difficult to open up particular datasets because issues such as the confi-

dentiality, data quality or the privacy infringement risks need to be addressed. Besides the privacy infringement risk, there might be other risks associated with the publication of OGD, such as publication of data against the law or possible misinterpretation of the data (Kucera & Chlapek, 2014). Ubaldi (2013a) points out that there are not only technical and legal challenges associated with the OGD initiatives but there are also challenges related to policy, financing, organization and culture. Chapter 4 provides a comprehensive overview of the organizational issues for opening up government data.

The abovementioned challenges and risks show that there is a need for an OGD publication methodology that would provide the responsible persons (publishers) with a clear guidance on how the OGD initiatives should be implemented and how the known challenges and risks should be addressed. If the challenges are not properly tackled it might prevent the expected benefits from being reaped (Ubaldi, 2013a). On the other hand, open data initiatives and practices take place in many different sectors, while users of open data often combine data from various domains.

In terms of the MePOD-VS methodology (Kucera, 2015) an Open Data initiative is an initiative executed by public sector bodies. Open Data publishing initiation might involve support of the top management of the public sector, and guarantee of departments and other stakeholders' participation. This is aligned with the SHARE-PSI 2.0 (2016) best practice on the *"Development of a Cross agency Strategy"*, which is presented in more detail in Chap. 4. According to (Moller, 2013), Open Data publication planning, Preparation of datasets and infrastructure, Open Data publication, cataloguing and maintenance and the Open Data archiving and retirement domains provide the necessary processes involved in the stages of the datasets lifecycle. Figure 2.2 illustrates the overall methodology and its process domains.

The main objective of the Open Data publication planning is to select a set of datasets for publication that is in line with the defined goals. The development of an open data publication plan will be used to steer the OGD initiative and it is aligned with the SHARE-PSI 2.0 (2016) best practice *"Open Data Publication Plan Development"*. Datasets planned to be released need to be prepared, e.g. they might need to be transformed into a suitable machine-readable format, enriched with metadata and properly licensed. Once the datasets are prepared they need to be made accessible and discoverable. Datasets and the respective metadata also need to be regularly updated (Lee, Cyganiak, & Decker, 2014).

Moreover, changes in legislation might affect what datasets particular public-sector organizations are able to publish as OGD, since the data could be characterized as private at some point after the beginning of the open data initiative. The Open Data archiving and retirement is part of the publication methodology in order to properly manage the end of the dataset lifecycle. Zuiderwijk et al. (2012a) have defined a process of selecting the data for publication. They argue that dealing with privacy-sensitive data, deletion policies, publishing after embargo periods instead of not publishing at all, adding related documents and adding information about the quality and completeness of datasets. The institutional context should be taken into account when using the guidance, as opening data requires considerable changes of organizations. Since the progress and impact evaluation of an OGD initiative is

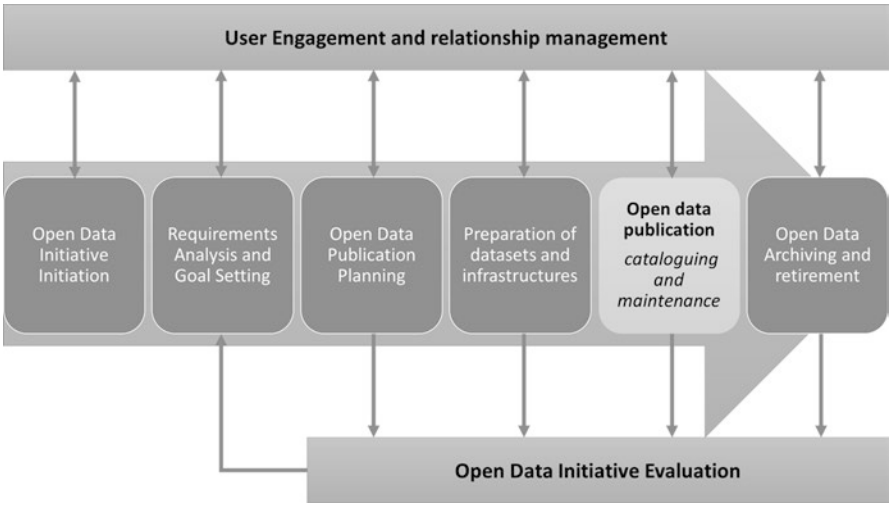

Fig. 2.2 Open data publication methodology, captured by Kucera (2015)

crucial for its development and implementation, a separate process domain is included dealing with the evaluation of progress against the Open Data publication plan and the defined goals.

User engagement and relationship management process domain is aimed at the identification of both actual and potential users of published data, the assessment of user's demands and requirements, as well as the setting up and execution of the communication strategy. It is also aiming at the assurance of feedback provision on the published data. While facilitation of the user feedback and re-use remains an important part of the OGD initiative this shift allows engaging users in the early stages of the OGD initiative which should help to establish a demand-driven release of data. This in turn should lead to a better alignment of data demand and supply.

Besides the tasks of the domains depicted in Fig. 2.2 there are other activities that need to be performed during the OGD publication such as the data quality management, benefits management or risk management (Nečaský et al., 2014). These topics are included as individual processes and not separate process domains. Since risk management and data quality should represent a continuous process, it is related to all process domains proposed in Fig. 2.2 in a way similar to the user engagement and relationship management process domain.

2.4.2 Towards Big Data Re-use: The Users' Side

Figure 2.3 presents a typical process of handling and processing big data in an enterprise environment beginning from the data identification towards data visualisation and utilisation of results.

Fig. 2.3 Big data user process adapted by Erl et al. (2016)

In a business environment the process starts with the identification of the problem to be tackled and the Key Performance Indicators (KPIs) that have to be measured determining the assessment criteria and guidance to the evaluation of analysis results. The problem to be solved should be quantified as a big data problem through the establishment of direct relations to one or more of the Big Data characteristics of volume, velocity, or variety. In Table 2.4 we describe the process step by step (Erl et al., 2016) and provide remarks on difficulties and crotchetiness for each one of them (Jagadish et al., 2014). Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportunities to utilize the analysis results. After Data Visualization stage, it might be needed to determine how and where processed analysis data can be further leveraged. Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce "models" that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.

2.4.3 Preparing a Scientific Data Infrastructure: Research Institutions

This subsection presents the user's perspective of the open data life cycle. As a user we have selected the researcher stakeholder. The constructors of the model begin with the statement that "Once the data is published, it is essential to allow other scientists to be able to validate and reproduce the data that they are interested in, and possibly contribute with new results" (Demchenko et al., 2013). Koop et al. (2011) argues that scientific data provenance should be taken into consideration by scientific data infrastructure providers.

Another aspect to take into consideration is to guarantee reusability of published data within the scientific community. Understanding semantic of the published data becomes an important issue to allow for reusability, and this had been traditionally being done manually. However, as we anticipate unprecedented scale of published data that will be generated in Big Data Science, attaching clear data semantic becomes a necessary condition for efficient reuse of published data. Learning from best practices in semantic web community on how to provide a reusable published data, will be one of consideration that will be addressed by the scientific data infrastructure. Big data are typically distributed both on the collection side and on the processing/access side: data need to be collected (sometimes in a time sensitive way

Table 2.4 Big data analysis process

Step	Description and remarks
Data Identification	Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for. Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise. In the latter case, open data could be found from third-party data providers, such as data markets and publicly available datasets, are compiled. Some forms of open data may be embedded within blogs or other types of content-based web sites, in which case they may need to be harvested via automated tools.
Data Acquisition and Filtering	Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter. In many cases, especially concerning external, unstructured data, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process. Since the data filtered out for one analysis may possibly be valuable for a different type of analysis, it is advisable to store a copy of the original dataset before proceeding with the filtering. To improve the classification and querying, metadata (e.g. dataset size and structure, source information, date and time of creation or collection and language-specific information) can be added automatically from both internal and external data sources. It is vital that metadata be machine-readable and passed forward along subsequent analysis stages. This helps to maintain data provenance throughout the Big Data analytics lifecycle, which helps to establish and preserve data accuracy and quality.
Data Extraction	This step realizes the extraction of data from the sources according to the filtering criteria of the previous step. The required extent of extraction and transformation depends on the types of analytics and capabilities of the Big Data tool (i.e. extracting text for text analytics, which requires scans of whole documents, is simplified if the underlying Big Data tool can directly read the document in its native format).
Data Validation and Cleansing	Since invalid data can skew and falsify analysis results, this an important step of the process. Big Data can be unstructured without any indication of validity. Most data sources are notoriously unreliable: sensors can be faulty, humans may provide biased opinions, remote websites might be stale, and so on. Its complexity can further make it difficult to arrive at a set of suitable validation constraints. Understanding and modelling these sources of error is a first step toward developing data cleaning techniques. Provenance can play an important role in determining the accuracy and quality of questionable data.
Data Aggregation and Representation	This step deals with the required data reconciliation method to determine and represent the correct value. Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. The large volumes processed by Big Data tools can make data aggregation a time and effort-intensive operation. Future data analysis requirements need to be considered during this stage to help foster data reusability. A standardised data structure could act as a common denominator that may be used for a range of analysis techniques and projects. This can require establishing a central, standard analysis repository, such as a NoSQL database.

(continued)

Table 2.4 (continued)

Step	Description and remarks
Data Modelling and Analysis	The data analysis step is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This step can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered. Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related, and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. In fact, with suitable statistical care, one can use approximate analyses to get good results without being overwhelmed by the volume.
Data Visualization	The last step of the process is to produce recognizable and useful insights through visuals to increase the value of the analysis of big data. The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users. Users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback or make the right decisions. The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. The same results may be presented in a number of different ways, which can influence the interpretation of the results. Consequently, it is important to use the most suitable visualization technique by keeping the business domain in context. Another aspect to keep in mind is that providing a method of drilling down to comparatively simple statistics is crucial, in order for users to understand how the rolled up or aggregated results were generated.

or with other environmental attributes), distributed and/or replicated. Linking distributed data is one of the problems to be addressed by SDI. The required new approach to data management and handling in e-Science is reflected in the Scientific Data Lifecycle Management in Fig. 2.4, as a result of analysis of the existing practices in different scientific communities.

The generic scientific data lifecycle includes several consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding). The Scientific Data Lifecycle Management necessitates data storage and preservation at all stages what should allow data re-use and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in scientific data infrastructure. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the scientific data lifecycle and must also be done in a secure and trustworthy way.

This example of scientific open data life cycle was selected based on its increased complexity compared to the two previous ones. The previous stakeholders do not

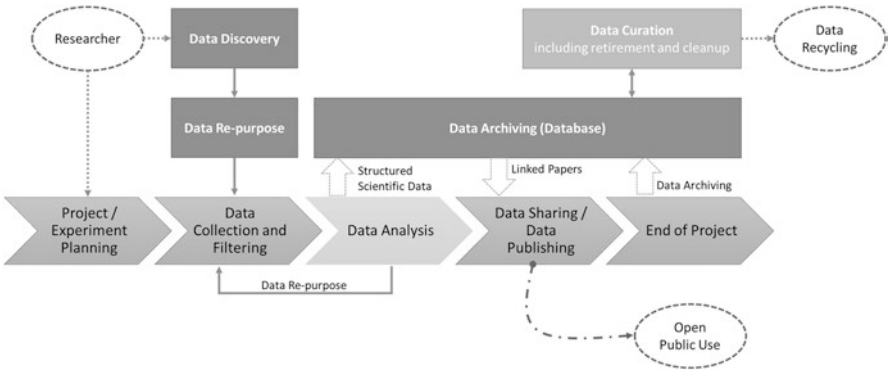

Fig. 2.4 Scientific data lifecycle management in e-science adapted from Demchenko et al. (2013)

pose so sophisticated requirements. Two are the most important issues regarding the peculiarities of this use case that are addressed by the open data life cycle model. Firstly, the recognition of the duality of a user to be both a user and a producer of data and secondly, the identification of the essential element of collaboration and interaction between different communities of users as well as between users and producers of data providing the necessary tools and workflows in the open data life cycle. These workflows will support the demand side of open data enhancing the exploitation step and closing the feedback loop.

2.4.4 Towards Linked Data Re-use: Publishers and Users

In order to support the full life cycle of linked open data, the Open Data Support Working Group resulted in the linked open data life cycle model presented in Fig. 2.5 including steps for both supply and demand (publishers and users) connecting them through the feedback step and thus closing the feedback loop.

In addition, the LOD2 stack is an integrated distribution of aligned tools which support the lifecycle of Linked (Open) Data from extraction to visualization and maintenance. The stack comprises tools from the LOD2 partners and third parties. With the ambition to identify these tools to support the creation and use of linked data, LOD2 project developed a more fine-grained 8-step life cycle model (Auer et al., 2012) formulated as follows: Extraction; Storing and Querying; Manual Revision and Authoring; Interlinking and Fusing; Classification and Enrichment; Quality Analysis; Evolution and Repair; Search and Browsing. Furthermore, LOD2 project has developed techniques for assessing quality based on characteristics such as provenance, context, coverage or structure. The open data life cycle presented in Sect. 2.3 has integrated these steps and tools incorporating the representation of linked data in the model, but this is not always the case. The LOD2 stack would guide better the manipulation of linked data since it is conceptualized and implemented targeting linked data specific characteristics. These specific characteristics towards data interoperability are mentioned and highlighted in Chap. 5.

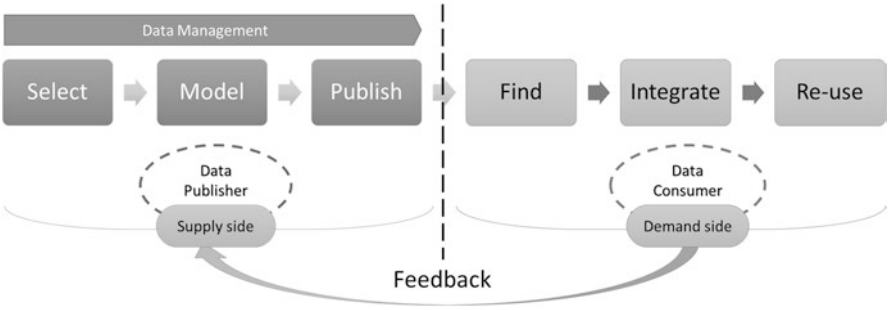

Fig. 2.5 OGD life cycle adapted from Open Data Support Working Group (https://joinup.ec.europa.eu/sites/default/files/D2.1.1%20Training%20Module%202.1%20The%20Linked%20Open%20Government%20Data%20Lifecycle_v0.11_EN.pdf)

2.5 Conclusions and Open Data Principles

This chapter identified the major data management and open data life-cycle models that exist in contemporary scientific literature. The major models have been presented in detail for each sub-category of technologies (linked data, big data) and associated stakeholders (publishers, users). Each life-cycle model could be used efficiently in different contexts. Finally, we introduced the new paradigm of the open data life cycle model from an ecosystem perspective including collaboration and feedback capabilities and acquainting with the notion of "data pro-sumer". A user with a possible dual role in the open data system being both producer and consumer of data.

The data itself is often treated as "a commodity rather than an artefact" (Meijer et al., 2014). However, how (open) data is understood and interpreted is shaped by the institutional and legal context, e.g. different perceptions of privacy and personal data. In a similar manner, some data can be considered more politicized than other. Also, different professional perspectives on data that refers to the same material object influence not only the sense-making, but the consideration of what data is actually important, the metrics of measurement etc. Altogether, this might even question the viability of a generic life-cycle model. Regarding the latter observation there should be an individual life-cycle model, which fits best in each situation.

Furthermore, this chapter identifies some principals for the open data that should be accompanying open data publication throughout its life-cycle. The principals for the open data publication process are:

Transparency-by-design (Janssen, 2015) Transparency-by-design refers to a principle where data about the functioning of government is automatically opened, can be easily accessed and interpreted, without being manipulated or being predefined or pre-processed. Transparency-by-design should ensure that information for effective public oversight is made available and that this information is clear and not ambiguous. Adherence to this principle requires that the mechanisms for

creating transparency are integrated in the heart of the government functions. This does not necessarily imply that all data is opened, but that all data necessary for effective oversight are open.

Privacy-by-design (Janssen, 2015) Privacy-by-design means that systems and the governance of these systems, are developed to guarantee individual privacy. Privacy-by-design does not mean that data cannot be shared. Privacy-by-design should also contain measures to compromise privacy for the sake of national security. Peled (2014) argues that restrictions such as authorization from individuals before their medical data are released are required to increase data circulation. Although the need for privacy and transparency is intuitively clear, realizing both principles is a complex endeavour that might be one of the thorniest problems in digital government. Transparency and privacy are inter-dependent and non-dichotomous variables and complete transparency and privacy does not exist. Both principles compete with each other as well as with other principles underpinning our society and individual versus collective rights and responsibilities. Weighing transparency versus privacy requires a deep understanding of the situation at hand.

Quality-by-design The quality of data could be seen and assessed from different perspectives. The basic data quality measurements are: accuracy, completeness, consistency and timeliness. Even more perspectives could be included in the quality assessment, such as comprehensiveness, speed, security, correctness and others that will fully analysed in Chap. 8: Open Data Evaluation. Except the standard quality measures, data quality is heavily connected with the metadata provision, as well as the ascription of a persistent URI ensuring the unique identification of an open dataset. Furthermore, Tim Berners Lee introduces the 5-stars open data maturity model for quality measurement towards linked data focused mainly on the format of the provided data.

Closing the feedback loop One essential element of open data ecosystems concerns their development "through user adaptation, feedback loops and dynamic supplier and user interactions and other interacting factors" (Zuiderwijk et al., 2014). Open data ecosystems perform data production and usage-cycles with feedback loops, sharing of data back to publishers and also with the so-called infomediaries (Pollock, 2011). However, discussion and feedback loops appear barely to be part of existing open data practices and infrastructures. Zuiderwijk and Janssen (2013) found that after open data have been used, the provision of feedback to data providers or a discussion with them is quite important by not facilitated by existing open data infrastructures, though such mechanisms might be useful for improving open data quality, data release processes and policies. Dawes and Helbig (2010) found that such mechanisms can help users to obtain insight in how they can use and interpret open government data and generate value from them.

Besides generic policies and concepts on open data (Directive 2003/98/EC on the reuse of public sector information and the European Data Portal), various other – thematic – policies and concepts determine, guide or influence the provision, and the use of open data. In some domains the process towards openness is supported by legislative EU frameworks. In the geospatial / environmental data domain there are: (a) the INSPIRE framework Directive 2007/2/EC, (b) the Directive 2003/4 on public access to environmental information and (c) the earth observation with the EU Regulation 1159/2013 on the European Earth monitoring programme (GMES). In the transport domain there is the Directive 2010/40/EU on the deployment of Intelligent Transport Systems in the field of road transport. There is also a data model for statistical information (SDMX: the Statistical Data and Metadata eXchange) and a data model for social sciences study-level information (DDI - Data Documentation Initiative). In addition, in other domains – and across domains – initiatives have been taken and actions have been setup to support and enable open data. For some domains, this is strongly based on a national responsibility to promote transparency of government processes and products (e.g., access to legal data such as legislation, jurisprudence through national records acts). Particular effort has been made to promote and facilitate the opening of research and education data (e.g., European Commission 2016).

Best practices for open data have been defined and assigned to each element of PSI Directive on the re-use of open data from the SHARE-PSI 2.0[1] EU project and some more technical ones from the Data on the Web Best Practices Working Group (2017) of W3C. The next chapters will introduce the concept of open data analysed from technological business, socio-technical, operational, process, legal and governance perspectives, while the open data ecosystem will be largely described by its individual elements.

[1] https://www.w3.org/2013/share-psi/bp/