FASED: A Framework for Data Ecosystems Health Evaluation

Glória de Fátima B. Lima[1], Marcelo Iury S. Oliveira[2], Bernadette Farias Lóscio[1]

[1] Universidade Federal de Pernambuco, Brazil
{gfbl, bfl}@cin.ufpe.br
[2] Universidade Federal da Paraíba, Brazil
{marcelo}@ci.ufpb.br

Abstract. The growing availability of data in digital media has contributed to the creation of a large number of data ecosystems. However, having successful Data Ecosystem is still a challenge. In order to prevent the failure of a Data Ecosystem and ensure its survival, evaluating its health becomes fundamental. In a general way, the health of a Data Ecosystem can be defined as its ability to grow and survive over time. Indicators such as productivity, robustness, niche creation and sustainability can be employed to evaluate the health of a Data Ecosystem. In this paper, we propose a framework for data Ecosystem health evaluation composed of a set of indicators and metrics, which assess the Data Ecosystem's current state and its ability to stay healthy over time. The results obtained when using the proposed framework offers evidence to assist in decision making on how data has being published and consumed in a Data Ecosystem, as well as to evaluate which ecosystems are more prosperous or need more investments.

Categories and Subject Descriptors: E. [**Data**]: Miscellaneous; C. [**Computer Systems Organizations**]: Miscellaneous; H.5 [**Information Storage and Retrieval**]: Group and Organization Interfaces

Keywords: Data Ecosystem, Data Ecosystem Health, Quality

1. INTRODUCTION

Governments, research institutions and individuals are producing and making large amounts of data available on various types of platforms (*e.g.* on the Web, applications applied to sensors and social media) [Chen et al. 2014; Silva et al. 2015]. According to [Pollock 2011], in the majority of these cases, the current basic model for the provision and usage of data is a one-way street. There is no feedback loop between data users and data consumers; *i.e.*, data users do not share data and knowledge back to their data producers. In order to unlock the potential benefits of sharing data, a Data Ecosystem needs to be established [Ubaldi 2013].

A Data Ecosystem (DAECO) can be seen as *"a network of actors composed of autonomous actors who directly or indirectly consume, produce or provide data and other data-related resources (e.g. software, services and infrastructure). Each actor plays one or more roles and is connected to other actors through relationships, so that collaboration and competition between the actors promotes the self-regulation of the Data Ecosystem"* [Oliveira and Lóscio 2018]. Indeed, various data from different organizations are used, re-used and exploited in cross-industry, socio-technical networks –so-called Data Ecosystems [Gelhaar et al. 2021; Oliveira and Lóscio 2018]. Some authors advocate that the engagement in ecosystems is no longer a choice, but it is mandatory for companies to unlock the benefits of data sharing [Gelhaar et al. 2021; Thomas and Autio 2014; Oliveira and Lóscio 2018]

An example of DAECO is the consumption and production of data on the Twitter platform. Twitter is a social media platform free to use for individuals and businesses alike. Most of Twitter revenue comes from advertising. But, Twitter established a Data Ecosystem that enables its several actors to

access and analyze historical and real-time data on the company's platform. It starts with politicians, celebrities and several common people publishing their thoughts, interacting, sharing content, and reading breaking news via posts-called tweets. These twitter users act as data producers producing a great amount of data (tweets) on a daily basis. Twitter itself acts as a data hub collecting data directly from its users. In its turn, the large scale of twitter data have tremendous usages by various data consumers. For example, companies use tweets to study customer behavior, monitor public responses to their products; public policy makers explore twitter data to obtain the demographic information for making strategic decisions; and sociologists leverage social media data to study social behavior and establish new social network theories [Zhang et al. 2018]. Meanwhile, there is a thriving myriad of Twitter developers building products and services that intermediate or ease data consumption. Companies, such as Gnip, Dataminr, DataSift and Topsy, help to make a large number of public tweets available to partners to build products and services as well as help other companies to get closer to customers on Twitter.

At the example above, Twitter acts as a keystone enabling the Data Ecosystem. However, other structural organizations are possible. The whole network of relationships may follow an organizational structure, ranging from an *ad-hoc* diffuse approach to a marketplace approach [Hanssen and Dybå 2012]. For example, Azure Data Marketplace, Snownflake Data Marketplace and IOTA Data Marketplace are examples of initiatives that provides data scientists, business intelligence and analytics professionals, data scientists, and others data consumers live access to ready-to-query data from data service providers. These kinds of data ecosystems are structured to facilitate the data trading and exchange process removing barriers that prevent providers and consumers from performing their activities. DAECO not only breaks organizational boundaries as internal data is increasingly used externally and vice versa [Lis and Otto 2020]. It also requires finding a balance between the conflicting interests of having control over data assets and willing to share data to design and deliver common value propositions [Lis and Otto 2020].

While DAECOs are gaining importance for their potential, some are still unable to remain productive, bring in new investments, deliver value and, therefore, do not survive for long periods of time. Consequently, the effort invested by their actors ends up not being well used or even forgotten. As a way of identifying deficiencies in functioning and meeting the expectations of the participants, it is necessary to have indicators that reflect the health of the DAECOs.

DAECO are not unlike biological ecosystems. They consist of multiple actors (*i.e.*, species) performing different roles, which need each other for success and survival. Again, like a biological ecosystem, a DAECO and its actors are also more or less "health". Health is a term in Biology, which refers to the status of the system or specific species [den Hartigh et al. 2006]. Iansiti and Levien (2002) have introduced the "health" as an overall performance indicator of business ecosystems. Since then, it have been adapted by other domains such as Software and Service to refer information about the current state of an ecosystem and its components as well as their ability to grow and survive over time.

However, the measurement of a DAECO health is not yet fully achieved. Instead, maintainers and keystone actors of DAECOs need to understand and make decisions about the socio-technical impact of technological, political and normative changes that affect the ecosystem health and recommend corrective actions. Unfortunately, there is little support or best practices for enabling DAECO maintainers to perform these tasks.

Assessment frameworks for validating the health of Data Ecosystems should provide the means to evaluate the functionality and status of elements in a DAECO. In this context, this work proposes a framework for assessing a DAECO health, called FASED, consisting of a set of indicators, characteristics and metrics. The proposed framework evaluates the main constructs (*e.g.* actors, relationships, resources and roles) of a DAECO [Oliveira and Lóscio 2018] and offers evidences about their health. Considering the lack of work in this area, we used as inspiration other ecosystem health assessments

frameworks such as [Iansiti and Levien 2002].

This work is organized as follows: Section 2 and 3 presents a theoretical background and the related works, respectively. Section 4 describes the research methodology, Section 5 presents our proposal, Section 6 details the assessment of the framework using the Focus Group method and Section 7 discusses the next steps of this research.

2. THEORETICAL BACKGROUND

Data Ecosystem field is inspired by the notion of Biological Ecosystems, which, in particular, denote the interactions between organisms and their environment as an integrated system [Chapin III et al. 2011]. A Data Ecosystem can be viewed as another instance of a Business Ecosystem, a Digital Business Ecosystem or a Software Ecosystem. Indeed, it borrows aspects from former ecosystems. However, Data Ecosystem does not rely on an explicit common platform in which different actors can collaborate. The common platform is actually the wide collection of datasets exchanged by the actors. In particular, data do not necessarily need to be provided by a single actor. The lack of a common platform creates a more diffused supply-demand network. Another difference is related to the perception of products traded between the actors. In Business Ecosystems, business operations and actors *per se* are the products [Manikas and Hansen 2013]. In Software Ecosystems, the products are software components or services. In Data Ecosystem, the products are data and their related technologies.

In this sense, Data Ecosystem can be envisioned as part of multiple types of ecosystems organized around businesses, resources and products provided by different actors. The broader goals of innovation and value creation are translated into more specific terms related to each particular ecosystem context. In particular, data can be used to support business, to deliver innovation, to promote transparency for governments, to validate research and numerous other goals.

The emergence of Data Ecosystems has been driven by several factors, including the emergence of digital technologies and political/institutional initiatives. For instance, most Data Ecosystems have been mainly driven by the Open Data movement, which call for the free use, reuse and redistribution of data by anyone [Group]. Several governments have already launched Open Data Portals to stimulate and promote Open Data production and consumption [Chun et al. 2010].

Typically, Data Ecosystems rely on a vast and heterogeneous set of actors, each one with different properties, capabilities and expectations. Similarly, Data Ecosystem resources are heterogeneous. For instance, datasets are heterogeneous regarding structural (schema), syntactic (format) and semantic (meaning) issues. Actors may produce and consume data using different activities and under different conditions. [Oliveira and Lóscio 2018; Oliveira et al. 2018] point that four main constructs stand out from DAECO literature: (1) actors, (2) roles, (3) relationships and (4) resources.

An actor is an autonomous entity such as an enterprise, institution or individual, which plays one or more specific roles in a Data [Oliveira et al. 2018]. Actors bound to a role must possess the capability of discharging the commitments a role imposes for them. A role is a function played by an actor in a Data Ecosystem [Oliveira et al. 2018]. It is related to a set of duties and activities. Relationships are the interactions among Data Ecosystem actors [Oliveira et al. 2018]. Relationships are often based on a common interest or are also related to the role each actor serves in the ecosystem. Actors exchange data or another type of resources through a transaction in a relationship. Finaly, resources are a useful or valuable product, possession or capability produced, provided, curated or consumed by Actors. In Data Ecosystems, resources range from datasets and data-based software to infrastructure [Oliveira et al. 2018].

3. RELATED WORKS

[Costanza 1992] defines the health of an ecosystem as its ability to maintain structure (organization) and function (vigor) over time even when facing external forces that generate stress (resilience). So far, alternatives to evaluate the health of Data Ecosystem are still naive, focusing on relatively simplistic metrics, such as number of published data sets, number and percentage of downloaded datasets, number of datasets scheduled for launch, number of APIs and basic site analysis (e.g., number of page views, downloads, etc.) [Dawes et al. 2016]. To the best of our knowledge, there are no studies published focused on evaluating the health of DAECOs [Oliveira et al. 2019]. However, other ecosystem domains have more mature research related to the topic.

Iansiti and Leviam (2002) introduced "health" as a general performance indicator for business ecosystems. According to them, the indicators of the health of the business ecosystem are:

—robustness, the ability of an ecosystem to face and protect from an ecosystem;
—productivity, the efficiency with which an ecosystem converts inputs into outputs;
—niche creation, the ability to create significant diversity and thus new capabilities.

In addition to defining productivity, robustness and niche creation, Iansiti and Levien (2002) listed the factors that determine these indicators, such as Total factor productivity, Productivity improvements, Delivery of innovations, Survival rates, Persistence of structure and Variety. However, none of the indicators and factors are defined in terms of operational metrics.

Iansiti and Levien (2002) is fundamental for outlining the ecosystem health research. They were the first providing guidelines on how ecosystem health may be operationalized. However, den Hartigh et al. (2013) were who first attempt to operationalize the health of a business ecosystem, based on the categories of health indicators presented by Iansiti and Levien (2002). The work presented by [den Hartigh et al. 2006] proposes a new definition for Business Ecosystems health and an instrument with operational metrics that can be used by managers to evaluate the health of Business Ecosystems.

The frameworks presented by Iansiti and Levien (2002) and Hartigh et al. (2013) have been inspirational for Software Ecosystems also. For instance, the work presented by [Jansen 2014] aims to define and analyze the functioning of Open Source Software Ecosystems, based on the Iansiti and Levien (2002) indicators: productivity, robustness and niche creation. The author suggests that the health of a Software Ecosystem can be defined by two factors: longevity and capacity for growth. Another observation made by the author is the difference between the health of software projects and the health of Software Ecosystems. The health of software projects is evaluated using metrics such as: tracking and correcting errors, number of releases and number of downloads. And the definition of Software Ecosystems health, on the other hand, depends on factors such as: connection between actors and the capacity to grow of Software Ecosystems. The author presents these factors as being complex to measure.

Alternatively, [Dhungana et al. 2010] presents a comparison between Natural and Software Ecosystems, and analyzes aspects of sustainability of Natural Ecosystems that are relevant to Software Ecosystems. In particular, it focuses on the challenge of how to bring up ecosystem sustainability factors, as this problem is common on both ecosystems. For this reason, identifying factors that promote the sustainability of the ecosystem involves identifying different aspects that must be met without imposing excessive control. The author cites some metrics in the literature, such as number of emails, commits, bug fixes, but he does not propose new ways of measuring sustainability or the implementation of metrics to evaluate health.

Another perspective was presented by [Franco-Bedoya et al. 2014], who proposes a model, called QuESo, built from quality aspects. This model was composed of two types of elements: quality characteristics and metrics. The quality characteristics correspond to the software attributes relevant

to the evaluation, organized in a hierarchy composed at the highest level by 3 dimensions. The dimensions are the Software Ecosystems platform, the community and its network of actors. The dimensions are divided into characteristics and these, in turn, are divided into sub-characteristics. Some features were extracted from the quality model for Open Source software, called QualOSS [Soto and Ciolkowski 2009]. The QuESo is able to analyze the quality of Software Ecosystems, however it does not address the influence of platform quality on the quality of products, present in Software Ecosystems. In addition, the model and metrics presented, as well as their indicators and characteristics, are presented in a conceptual way, without any form of calculation and interpretation.

In its turn, [Carvalho et al. 2017] proposes a three-tier architecture in order to assess the health of Software Ecosystems, called HEAL ME. The architecture is composed by the communication, service and knowledge discovery layers. As a way of evaluating health, HEAL ME uses the indicators proposed in the works of [Dhungana et al. 2010] and [Iansiti and Levien 2002] to measure the health of Software Ecosystems. In addition to previous works, the authors also used the work [Jansen 2014] and the quality model proposed by [Franco-Bedoya et al. 2014] as sources to extract the metrics. [Carvalho et al. 2017] structured the health evaluation into a hierarchy of indicators, characteristics and metrics, where 5 indicators, 9 characteristics and 58 metrics can be found. And, finally, a case study was done in order to evaluate the viability and applicability of HEAL ME, in the context of a scientific Software Ecosystems.

Table I: Related Works Comparison

Study	Ecosystem Domain	Indicators	Multiple Levels	Metrics
[Iansiti and Levien 2002]	Business Ecosystem	Productivity, Robustness and Niche Creation	NO	NO
[den Hartigh et al. 2006]	Business Ecosystem	Productivity, Robustness and Niche Creation	Business Ecosystem Level and Company Level	YES
[Jansen 2014]	Software Ecosystem	Productivity, Robustness and Niche Creation.	Network level and Project Level	YES
[Dhungana et al. 2010]	Software Ecosystem	Sustainability	Technical Issues, Business Consideration, and Community Participation	NO
[Franco-Bedoya et al. 2014]	Software Ecosystem	Maintenance capacity, Process maturity, Sustainability, Network health, Resource health	Platform, Community and Ecosystem Network level	YES
[Carvalho et al. 2017]	Software Ecosystem	Sustainability, Diversity, Productivity, Robustness, Niche Creation	NO	YES

These works above show the importance of health assessment and the impact of this assessment on the functioning of ecosystem components. Table I presents a summary of the works presented. As can be seen, the work of Iansitti and Levien (2002) has influenced evaluation in proposing new frameworks and metrics for health assessment. Some works, in addition to indicators, recognize that the measurement of health in an ecosystem must be carried out in a multi-level perspective. Each level may have different metrics, and corrective actions are different depending on the level. In addition to the indicators proposed by Iansitti and Levien (2002), the dimensions Sustainability and Diversity were proposed. In another perspective, the work carried out proposes to evaluate Software Ecosystems by technical aspects related to software development projects, for this reason, it chose to call Quality Assessment instead of ecosystem health assessment.

In addition, the work [den Hartigh et al. 2006] was a pioneer in the measurement of health, in terms of metrics operationalization. Since the 80s, with the advent of Total Quality Management, metrics have been used to help monitoring and managing organizations and systems. A common quote about performance measurement states "if it cannot be measured, it cannot be managed". Metrics are used to drive improvements and help businesses focus on what is important. A Data Ecosystem, as a

distributed, diffuse and heterogeneous organization, can benefit from metrics to improve management aspects.

This work is the first step in the area of Data Ecosystems health evaluation and aims to propose and explore generic metrics based on the components of DAECOs and related activities, to enable the evaluation of the ecosystem under study.

4. RESEARCH METHODOLOGY

As different research methodologies serve different purposes, one of the first steps is to choose a philosophical research paradigm that is appropriate [Easterbrook et al. 2008]. In this sense, the proposal of this paper is based on a pragmatic philosophical paradigm. According to [Easterbrook et al. 2008], pragmatism values practical knowledge about abstract knowledge and uses all appropriate methods to obtain it.

In addition, this research is also based on the Design Science Research paradigm, which aims to build relevant artifacts in terms of value and usefulness both at a practical and theoretical level [Hevner and Chatterjee 2010]. It is fundamentally a problem solving methodology that aims to create and evaluate IT artifacts that solve an important organizational problem.

The framework was built through an iterative and incremental process. As a result of a vast literature review on Data Ecosystems, a theoretical basis on Data Ecosystems has been consolidated. Still for the construction of the theoretical basis, studies were analyzed aimed at proposing frameworks for health evaluation and quality models. In particular, these works helped us to understand how they were built, from the study of the domain context to the evaluation, using specific metrics for each indicator. Both studies guided the construction of FASED.

The framework is influenced by ISO/IEC 25000: 2014 and some works identified during the ad-hoc study. The ISO/IEC 25000: 2014 (International Organization for Standardization and International Electrotechnical Commission 25000: 2014), which presents the SQuaRE (Systems and software Quality Requirements and Evaluation), brings together a series of international standards in metrics and quality models, as well as in evaluation and quality requirements for systems and software products. The choice of this standard was due to its generic nature, as it contains concepts of quality at a high level, enables the construction of hierarchies of quality characteristics and also because it is a standard widely disseminated in the literature. Among the works that influenced the framework, are the works of [Jansen 2014] and [Carvalho et al. 2017] in the area of Software Ecosystem, and the work of [den Hartigh et al. 2006] in the area of Business Ecosystems. These works propose health evaluation frameworks using a top-down approach that identifies health indicators as a starting point and refines to identify the metrics.

From the initial set of concepts, the FASED design was continuously checked and refined. The construction process, in which the refinement and verification phases were carried out cyclically, is inspired by the Twin Peaks [Nuseibeh 2001] model. The process is iterative and progressively produces more detailed requirements and specifications.

Thus, the construction started from the elaboration of an initial version of the framework based on the ad-hoc study of frameworks proposed in other types of ecosystems and in the quality models of software evaluation. This initial version has been improved through verification and refinement cycles. All elements that compose the FASED were checked for their contribution to the definition of the framework and for the evaluation of DAECOs health. After that, the results of the verification process were used to refine the proposed framework.

5. FASED: FRAMEWORK FOR DATA ECOSYSTEMS HEALTH EVALUATION

The proposed framework was designed to support the health evaluation of Data Ecosystems. In general, a health framework can be applied by researchers or practitioners who wants to evaluate the health of one ecosystem over another, or identifying weaknesses in an ecosystem with the aim of making it healthier [Jansen 2014]. FASED is aimed at providing comprehensive instruments to determine the current health of a Data Ecosystem as well as its health over time. It does so by creating an inventory of four types of interrelated elements within a multi-level model.

The FASED development approach consisted of two main phases: creation of the theoretical basis and the framework development. In order to create a theoretical basis regarding Data Ecosystems and ecosystem health assessment, several studies were carried out, including, but not limited to, systematic reviews of the state-of-the-art on Data Ecosystems [Oliveira et al. 2019], and investigations about the characterization, definition, and modeling of Data Ecosystems [Oliveira and Lóscio 2018; Oliveira et al. 2018].

After the theoretical basis creation, the construction of the framework took place through an iterative and incremental process. Initially, the alpha version of the FASED framework was conceived based on influences from ISO/IEC 25000:2014 and from papers identified during the ad-hoc study. The choice of this standard was due to its generic nature, as it contains concepts of quality at a high level, enables the construction of hierarchies of quality characteristics and also because it is a standard widely disseminated in the literature. [Jansen 2014] and [Carvalho et al. 2017] in the area of Software Ecosystem influenced the framework creation as well as the work of [den Hartigh et al. 2006] in the area of Business Ecosystems. These papers propose health evaluation frameworks using a top-down approach that identifies health indicators as a starting point and refines them to identify the assessment metrics.

Beginning with the definition of the initial set of indicators, characteristics and metrics, the design of FASED was continually verified and refined. A complete analysis of the framework components were performed to check on how they could contribute to evaluate a Data Ecosystem health. After that, the results of the verification process were used for the refinement of the framework.

The FASED structure is organized in a multilevel hierarchy presented in Figure 1 and described in the following. Initially, the framework is defined by means of indicators, which in turn are decomposed into a group of characteristics. Next, the characteristics are refined considering different attributes, which are implemented by metrics. Such structure was inspired by standards, models, and frameworks investigated in the course of this work, such as ISO/IEC 25000:2014 and [Soto and Ciolkowski 2009; Franch and Carvallo 2003].

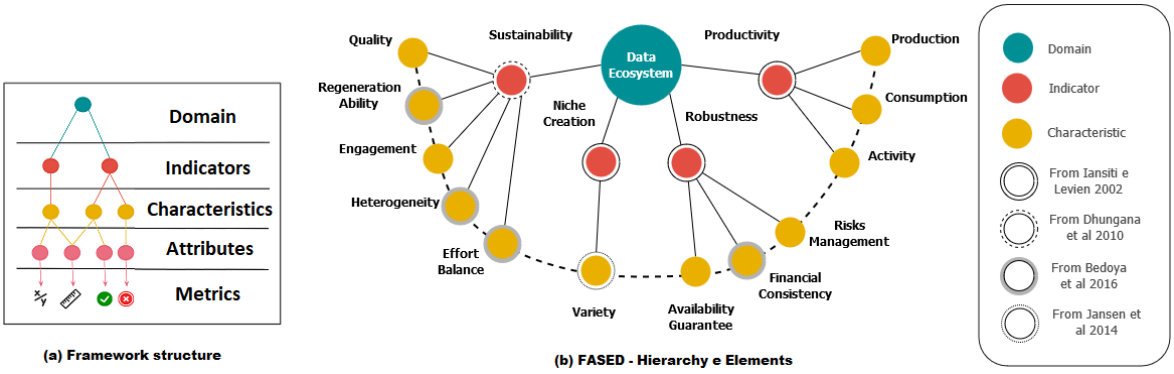

Fig. 1: FASED Framework

Indicators are defined to keep track of relevant concepts. Similar to [Jansen 2014], we use the three determinants and the underlying factors defined by Iansiti and Levien (2002) as a starting point, which are Productivity, Robustness, and Niche Creation. Beside the mentioned factors, we also considered a new indicator called Sustainability.

Then, once these indicators clearly establish the set of concepts to be tracked, we identify a set of **Characteristics** that will impact that indicator. Characteristics are non-measurable quality factors used with the aim of classification of the inventory of metrics.

Each characteristic is further refined according different **Attributes**, which represent forces that influence the health of a Data Ecosystem. We used the four basic elements of a Data Ecosystem identified by [Oliveira and Lóscio 2018]: actors, relationships, roles and resources. This design decision (*i.e.*, decompose characteristics into core elements of a ecosystem) is aligned to the one proposed by [Manikas and Hansen 2013].

In the final level, the characteristics and attributes are derived into **Metrics**, which are a quantifiable measure used to track and evaluate the status of a specific characteristic. While a metric evaluates a characteristic, an indicator will use one or more metrics to measure some topic. In order to properly interpret the metrics, each metrics is defined as the structure proposed by [Carvalho et al. 2017] and exemplified in Table II. Each metric has the following information:

(1) Indicator: to which the metric belongs;
(2) Metric name;
(3) Description: metric purpose;
(4) Measure and Formula: shows how the metric can be calculated;
(5) Interpretation: shows how the metric result can be interpreted;
(6) Unit: result measurement unit;
(7) Attribute: ecosystem construct related to the metric;
(8) Bibliographic Reference: if the metric was originally defined by another study, the reference for this scientific study will be presented in this field. Otherwise, it was a metric proposed originally to FASED, so the reference "Authors (2019)" is used.

It is important to remark that FASED follows a more descriptive than prescriptive approach. A prescriptive approach involves specifying, or even imposing, to individuals how they should do, rather than giving suggestions or describing what should be done. In its turn, the FASED focuses on presenting which factors should be considered by actors interested in evaluating a DAECO health. Such descriptive approach allows adapting to the reality of a Data Ecosystem.

Table II: Metric of Productivity Indicator to Evaluate DAECO's health

ID	Indicator	Metric	Description	Measure and Formula	Interpretation	Unit	Attribute	Bibliographic Reference
P1	Productivity	Volume of data produced	Volume of data produced by the DAECO and available for use	$X = \sum_1^T N_T$ N_T = Volume of data T = Period of time	$X >= P$ P = Parameter The greater the volume of data the better	Un - Unit	Resources	Author (2019)

In the next subsections, the indicators, their characteristics and metrics will be presented in more detail. For the lack of space, we have omitted some of the elements used to specify the metrics. The whole set of metrics with their full definitions is available at `https://repositorio.ufpe.br/handle/123456789/38286`.

Table III: List of Metrics

Productivity	**Characteristic: Production**
	(P1) Volume of data produced and available for use
	(P2) Volume of solutions produced and available for use
	Characteristic: Consumption
	(P3) Number of accesses to resources
	(P4) Number of downloads made to resources
	(P5) Number of external ecosystem solutions that consume data
	(P6) Ecosystems solutions that consume ecosystem data
	Characteristic: Activity
	(P7) Number of events that promote the ecosystem
	(P8) Number of reference materials to instruct the actors to use the ecosystem resources
Robustness	**Characteristic: Financial Consistency**
	(R1) Number of permanent private investors
	(R2) Number of permanent public investors
	(R3) Number of one-off private investors
	(R4) Number of public actors that contribute to the ecosystem financing in a timely manner
	(R5) Volume of financial resources invested in the ecosystem
	(R6) Volume of financial resources produced by the ecosystem
	(R7) Volume of financial resources spent by the ecosystem
	(R8) Number of different forms of monetization in the the ecosystem
	Characteristic: Availability Ensurance
	(R9) Existence of data backup on the ecosystem
	(R10) Existence of infrastructure that ensure 24/7 availability of resources
	(R11) Access by multiple mechanisms
	Characteristic: Risk Management
	(R12) Number of different types of organizational structures in the ecosystem
	(R13) Number of actors working in the management of the ecosystem
	(R14) Existence of audit in the ecosystem
	(R15) Existence of processes that support in the execution of the actors' activities
	(R16) Existence of documentation that describes the processes of the activities performed by the actors

Niche Creation	**Characteristic: Variety**
	(C1) Number of domains represented by the ecosystem data
	(C2) Number of different types of solutions produced by the ecosystem
	(C3) Number of different types of private organizations that are part of the ecosystem
	(C4) Number of different types of public organizations that are part of the ecosystem
	(C5) Number of different types of papers in the ecosystem
Sustainability	**Characteristic: Regeneration Ability**
	(S1) Active actors since the beginning of the ecosystem
	(S2) Number of new actors that entered the ecosystem
	(S3) Number of actors who leaved the ecosystem
	Characteristic: Effort Balance
	(S4) Ratio between the number of activities that only an actor or a small group performs and the total activities of the ecosystem
	(S5) Ratio between the number of actors who play a specific role and the total number of actors
	Characteristic: Heterogeneity
	(S6) Ratio between the number of partnerships in more than one country/state/city and the total number of the ecosystem partnerships
	Characteristic: Engagement
	(S7) Number of citations about ED on social media
	(S8) Existence of a communication channel between the consumer and the producer
	(S9) Participation of ecosystem actors in events from other ecosystems
	(S10) Number of works in the literature that mention the ecosystem
	Characteristic: Quality
	(S11) Data is easily accessed by downloading, API or other means of access
	(S12) Number of different formats in which data are published
	(S13) There is removal of data that is not accessed in a period of time
	(S14) License usage
	(S15) Protection against not authorized access
	(S16) Datasets with metadata
	(S17) Cleansing or refinement of data
	(S18) Data standardization
	(S19) Existence of mechanisms to check data veracity
	(S20) Data timeliness

5.1 Productivity

According to [Iansiti and Levien 2002], in Nature's Ecosystems, Productivity is defined as the effective capacity to convert raw materials into living organisms. Following this definition and considering that Business Ecosystems are constantly subject to new conditions, such as new technologies, processes and demands, [Iansiti and Levien 2002] define Productivity as the ability to convert raw materials of innovation into new products and functions at reduced cost. Taking as a starting point these definitions, we define Productivity as the ability of a DAECO to produce new resources (*e.g.* data, solutions and services) in order to allow the consumption of these resources and carry out activities that promote productivity.

The Productivity indicator has three characteristics, which are: Production, Consumption and Activity.

Production indicates the current state of resource production. The objective is to measure the volume of resources produced, as the quantification demonstrates the productive capacity of the ecosystem and, consequently, adds value to the ecosystem. **Consumption** indicates the current state of consumption of DAECO resources. The objective is to measure the volume of resources consumption that were produced by the ecosystem, as the quantification of this consumption stimulates the maintenance of existing resources and the production of new solutions. **Activity** indicates the existence of activities (*e.g.* hackathons, workshops, reference materials) that promote the production and consumption of resources by the DAECO actors.

These three characteristics group metrics to calculate two different aspects that influence health of a DAECO: the current state of productivity and the promotion of productivity.

Table IV: Metrics of Productivity Indicator

ID	Metric	Measure and Formula
Indicator: Productivity - Characteristic: Production		
P1	Volume of data produced	$X = \sum_1^T N_T$ N_T = Volume of data T = Period of time
P2	Volume of solutions produced	$X = \sum_1^T N_T$ N_T = Volume of solutions T = Period of time
Indicator: Productivity - Characteristic: Consumption		
P3	Number of accesses to resources	$X = \sum_1^T N_T$ N_T = Number of accesses T = Period of time
P4	Number of downloads made to resources	$X = \sum_1^T N_T$ N_T = Number of downloads T = Period of time
P5	Number of solutions from external ecosystems that consume the DAECO data	$X = \sum_1^T N_T$ N_T = Number of solutions from external ecosystem T = Period of time
P6	DAECO's solutions that consume their own data	X = N_i / N_t N_i = Number of solutions produced that consume the DAECO's data N_t = Total number of the DAECO's solutions
Indicator: Productivity - Characteristic: Activity		
P7	Number of events that promote the DAECO's resources	$X = \sum_1^T N_T$ N_T = Number of events T = Period of time
P8	Number of reference material to instruct the actors to use the DAECO's resources	X = N N = Number of reference materials

We have identified eight metrics to measure the Productivity of a DAECO (see Table IV). The assessment of these metrics is based on the capacity of the DAECO to produce new resources, which depends on the volume of produced resources; the capacity of consumption, through the numbers of accesses and downloads of data and solutions; and the existence of activities to promote productivity, such as producing events and documents to facilitate the use of resources.

5.2 Niche Creation

According to [Iansiti and Levien 2002], the Niche Creation in Nature's Ecosystems indicates the level of species variety as well as the support for species diversity. Similarly, [Iansiti and Levien 2002] defines Niche Creation in Business Ecosystems as the indication of the increasing of diversity over time through the creation of add-value functions. Based on these definitions, Niche Creation in a DAECO shows the variety of ecosystem elements in order to identify opportunities for the emergence of new niches.

The Niche Creation indicator has only one characteristic: **Variety**. It indicates the existence and capacity to promote the creation of niches by analyzing the variety of elements of the ecosystem, such as actors, roles and resources.

We have identified five metrics to measure the Niche Creation indicator (see Table V). In general, these metrics evaluate two aspects that influence health: the current state of niche creation and the ability to promote new niches over time.

Table V: Metrics of Indicator Niche Creation

ID	Metric	Measure and Formula
Indicator: Niche Creation - Characteristic: Variety		
C1	Number of domains represented by the data	X = N N = Number of domains represented by the data
C2	Types of solutions	X = N N = Number of types of solutions
C3	Types of private organizations	X = N N = Number of types of private organizations
C4	Types of public organizations	X = N N = Number of types of public organizations
C5	Types of roles	X = N N = Number of different types of roles

These two aspects are measured together in the Variety characteristic, as we can obtain information about the variety of domains represented by the data, as well as the types of resources produced, types of roles played by the actors and the types of public organizations (*e.g.* universities, development agencies) and private (*e.g.* hospitals, multinationals) that are part of DAECO. This information provides evidence of both the current state and the ability to create new niches.

5.3 Robustness

According to [Iansiti and Levien 2002], the Robustness indicator in Nature's Ecosystems is defined as the ability to persist in the face of changes in the environment. Similar to this definition, according to [Iansiti and Levien 2002], Business Ecosystems should be able to face and survive disruptions. Based on these definitions, we define Robustness of a DAECO as its ability to remain stable when facing disruptions. In this work, the term disruption means disturbance or problems that interrupt an event, activity or process [Oxford Dictionary 2019].

The Robustness indicator has three characteristics, which are: Financial Consistency, Availability Guarantee and Risk Management. **Financial Consistency** indicates the ability of a DAECO to keep up with the investments received and financial resources produced by it, while attracting new investors and partners [Franco Bedoya et al. 2016]. **Availability Guarantee** indicates if there are mechanisms that guarantee the continuous availability of resources (*i.e.* data, applications, services, infrastructure) even in situations of instability. **Risk Management** indicates the existence of actions and resources that assist in ecosystem management, such as auditing, well-defined organizational structures, actors and processes. They help prevent risks in the event of disturbances in the ecosystem.

These three characteristics group metrics that calculate two different aspects that influence DAECO's health: the current state of robustness and the ability to remain robust over time.

We have identified sixteen measures to assess the Robustness indicator (see Table VI). In particular, the current state of robustness is calculated by metrics that evaluate: i) whether the DAECO has the means to guarantee the continuous availability of resources in situations of instability, ii) financial consistency through the mapping of investors (*i.e.* punctual and permanent), iii) financial resources produced, received and spent by the DAECO, as well as iv) the variety of available data monetization strategies (*e.g.* sales of digital products, pay per use, sale of user information).

The ability to remain robust is also influenced by the presence or absence of risk management in the ecosystem. The existence of management in the ecosystem reveals the adoption of preventive measures

Table VI: Metrics of Robustness Indicator

ID	Metric	Measure and Formula
Indicator: Robustness - Characteristic: Financial Consistency		
R1	Number of permanent private investors	$X = \sum_1^T N_T$ T = Period of time N_T = Number of private investors that finance the ecosystem permanently
R2	Number of permanent public investors	$X = \sum_1^T N_T$ T = Period of time N_T = Number of public investors that finance the ecosystem permanently
R3	Number of temporary private investors	$X = \sum_1^T N_T$ T = Period of time N_T = Number of private investors that finance the ecosystem temporarily
R4	Number of temporary public investors	$X = \sum_1^T N_T$ T = Period of time N_T = Number of public investors that finance the ecosystem temporarily
R5	Volume of financial resources invested in the DAECO	$X = \sum_1^T N_T$ N_T = Volume of financial resources invested T = Period of time
R6	Volume of financial resources produced by the DAECO	$X = \sum_1^T N_T$ N_T = Volume of financial resources produced T = Period of time
R7	Volume of financial resources spent by the DAECO	$X = \sum_1^T N_T$ N_T = Volume of financial resources spent T = Period of time
R8	Monetization forms	X = N N = Number of different forms of monetization
Indicator: Robustness - Characteristic: Availability Guarantee		
R9	Data backup	X = B B = Existence of data backup
R10	Existence of infrastructure that guarantees resources availability 24/7 in the DAECO	X = B B = Existence of infrastructure that guarantees resources availability 24/7
R11	Access by multiple mechanisms	X = N_i / N_t N_i = Volume of data accessible by more than one mechanism N_t = Total volume of data
Indicator: Robustness - Characteristic: Risk Management		
R12	Number of different types of organizational structure existing in the DAECO	X = N N = Number of different types of organizational structure
R13	Number of actors that operate in the DAECO management	$X = \sum_1^T N_T$ N_T = Number of actors that operate in the management T = Period of time
R14	Existence of audit in the DAECO	X = B B = Existence of audit
R15	Existence of processes to perform activities	X = B B = Existence of processes to perform activities in the DAECO
R16	Existence of process documentation	X = B B = Existence of documentation

in case of disturbances in the functioning of the ecosystem. The presence of different organizational structures (*e.g.* Oriented to business models, centered on keystones, centered on platforms) increases the robustness, as they protect the actors against external disturbances.

In addition, the presence of well documented processes to guide the actors in the execution of activities contribute to risk management.

5.4 Sustainability

According to [Chapin III et al. 1996], Sustainability is one of the main challenges in any ecosystem. A sustainable Nature's Ecosystem maintains the diversity of the main functional groups, productivity and bio-chemical cycles, even in the face of events that disturb the natural state of that ecosystem. Similar to the definition of [Chapin III et al. 1996], [Dhungana et al. 2010] considers a software ecosystem as a sustainable one if it is able to increase or maintain its products, resources, members and relationships for long periods of time and manages to survive changes such as new technologies, new products and competitors. Similar to these definitions, we consider that DAECOs are sustainable when they have the capacity to increase or maintain their resources, actors and relationships for long periods of time and survive changes, such as new technologies, entry and exit of actors as well as market competitors.

The Sustainability indicator has five characteristics, which are: Regeneration Ability, Effort Balance, Heterogeneity, Engagement and Quality.

Regeneration Ability indicates the evolution of the network of actors, through the number of actors that enter and interrupt their participation in the ecosystem. This analysis shows whether the DAECO has needed to regenerate in the past, indicating that this ecosystem is more likely to survive in the future if there are significant losses of actors [Franco Bedoya et al. 2016].

Effort Balance indicates the level of efforts distribution in a DAECO. This characteristic is relevant because the greater the knowledge concentration in small groups, the greater the risk if these groups are not present in the ecosystem [Franco Bedoya et al. 2016]. **Heterogeneity** indicates the diversity of DAECO that influences its growth and survival [Franco Bedoya et al. 2016]. This characteristic can be identified by the diversity in the geographical distribution of DAECO.

Engagement provides information on stakeholder participation in a DAECO. The participation of the actors can happen through events that promote the DAECO, publications of resources that increase the visibility of the ecosystem or through the use of communication channels through feedback. The increase in engagement shows the interest of the actors in the ecosystem and attracts new collaborators to contribute and support the ecosystem. **Quality** indicates whether the data is fit for use by consumers. This characteristic gathers information about the quality of resources made available by the ecosystem, through the assessment of accessibility, consistency and reliability, thus influencing consumption and, consequently, the survival of the ecosystem.

Table VII: Metrics of Sustainability Indicator

ID	Metric	Measure and Formula
Indicator: Sustainability - Characteristic: Regeneration Ability		
S1	Active actors since the beginning of the DAECO	X = N_i / N_t N_i = Number of active actors that participate since the beginning of the ecosystem N_t = Number of active actors that are part of the ecosystem
S2	Entry of new actors in the DAECO	$X = \sum_1^T N_T$ N_T = Number of new actors that joined the DAECO T = Period of time
S3	Exit of actors from DAECO	$X = \sum_1^T N_T$ N_T = Number of actors that left the ecosystem T = Period of time
Indicator: Sustainability - Characteristic: Effort Balance		
S4	Centralization of efforts	X = N_i / N_t N_i = Number of activities performed by an actor N_t = Total number of activities in the DAECO
S5	Role Distribution	X = N_i / N_t N_i = Number of actors that play a specific role N_t = Total number of actors in the DAECO
Indicator: Sustainability - Characteristic: Heterogeneity		
S6	Decentralization of partnerships	X = N_i / N_t N_i = Number of partneships in more than one country/state/city N_t = Total number of partneships in the DAECO
Indicator: Sustainability - Characteristic: Engagement		
S7	Citation about the DAECO on social media	$X = \sum_1^T N_T$ N_T = Number of published citations T = Period of time
S8	Communication channel	X = B B = Existence of a communication channel between actors
S9	Events in external ecosystems	X = B B = Attendance on events of external ecosystems
S10	Literature citations	$X = \sum_1^T N_T$ N_T = Number of publications in the literature that cite the DAECO T = Period of time
Indicator: Sustainability - Characteristic: Quality		
S11	Ease of access	X = B B = If the data is ease of access
S12	Data formats	X = N N = Number of different formats
S13	Data removal	X = B_T B_T = If data not accessed in a period of time are removed
S14	License usage	X = B B = If the data is restricted by any license
S15	Protection against not authorized access	X = B B = If confidential/personal data are protected
S16	Datasets Metadata	X = B B = If the datasets have metadata
S17	Cleansing or refinement of data	X = B B = If inconsistent data are removed
S18	Data standardization	X = B B = If there is standardization on how the data are presented
S19	Existence of mechanisms to check data veracity	X = B B = If there are mechanisms to check the data veracity
S20	Data update	X = Na / Nt Na = Amount of data updated N_t = Total amount of data

We have identified twenty metrics that can be used to measure the indicator Sustainability (see Table VI).

The ability to remain sustainable can be assessed by measuring the balance of effort, as the likelihood of the ecosystem becoming unstable becomes high if an actor or a small group of actors that deals with a specific activity leaves the ecosystem. Heterogeneity is an important aspect because the geographical distribution of the partnerships reveals the ecosystem's ability to survive if other locations can no longer support it. Partnerships are a strategic part of the ecosystem, as they are one of the ways to motivate actors to contribute [Oliveira et al. 2018].

The metrics of participants engagement in the ecosystem reveal the participation and interactions in social media, which have a wide reach nowadays, as they have the power to attract and add more actors to the DAECO. In addition to citations on social media, engagement can be measured by the existence of communication channels between the actors, events promoting the DAECO in other ecosystems and citations in the literature, such as theses, articles, patents, among others [Botelho and de Oliveira 2015]. These metrics offer information about the degree of public interest and the potential growth of a DAECO.

Finally, metrics of quality evaluate the data produced by the ecosystem focusing on aspects influencing data consumption, such as security, accessibility and consistency.

6. FRAMEWORK EVALUATION

The evaluation of the FASED framework was conducted by an empirical study using the focus group method. According to [Kontio et al. 2004], a focus group is a technique that comprises a group of participants gathered to discuss a particular problem or assess a particular topic. Focus groups have become popular in several areas such as Medicine, Social Sciences, Biology and Information Sciences [Zaganelli et al. 2015].

For [Kontio et al. 2004] the focus group is a quick and economical method to obtain experiences from professionals and users, as it can provide qualitative data and rich information about the topic discussed. In addition, with the application of this method it is possible to reveal perceptions that are difficult or expensive to capture with other methods [Kontio et al. 2004]. Thus, because it has a more dynamic characteristic, the data generated by the focus groups tend to be valuable and deeper than those collected by one-to-one collection methods, such as interviews or surveys [Barry et al. 2008].

6.1 Focus Group Protocol

Our process for conducting the Focal Group followed the steps suggested by [Chiara 2005]. The main objective of the study was to carry out a high-level assessment regarding the feasibility, completeness and adequacy of the FASED. As first steps, we defined the criteria for selecting participants, decided the session length, designed the sequence of questions to ask during the session, and prepared documents to provide the participants with the study background and objectives.

In order to gain insights about improvements and gaps in the FASED, we used the following criteria for selecting the participants:

—knowledge and expertise in Data Ecosystems;
—knowledge of Ecosystem Health in any domain;
—willingness to share their experiences and candid opinion.

In addition, to ensure proper discussion and interaction during the session, another criterion was to invite participants who knew each other as friends or co-workers. According to these selection criteria, our study needed practitioners in Data Ecosystems and/or Ecosystem Health. Such practitioners are

usually very busy and are not likely to respond to invitations from unfamiliar sources. Thus, a random sampling was not viable. As matter of fact, recruiting participants is a significant challenge for any research project. Ten practitioners were considered as candidates for this study and were contacted. Two did not answer the invitation. Eight accepted the invitation. Participants were guaranteed anonymity and all data has been anonymized.

Table VIII: Focus Group Participants Characterization

Participant	**Current Position**	**Educational Degree**	**Data Ecosystem Experience**	**Current Position Domain**
Participant 1	Master Student/ IT Technician	Computing Certified	2 - 3 years	Industry and Academy
Participant 2	Master Student/ Data Engineer	B.Sc. in Computer Science	2 - 3 years	Industry and Academy
Participant 3	Master Student/ Researcher	B.Sc. in Computer Science	2 - 3 years	Academy
Participant 4	Master Student/ Software Engineer	B.Sc. in Computer Science	None	Industry and Academy
Participant 5	Master Student/ Data Analyst	B.Sc. in Computer Science	2 - 3 years	Industry and Academy
Participant 6	Ph.D. Student/ Professor	Master in Computing Science	None	Academy
Participant 7	Ph.D. Student/ System Analyst	Master in Computing Science	4 - 5 years	Industry and Academy
Participant 8	Master Student	B.Sc. in Computer Science	2 ¿ years	Academy

Table VIII presents the general characteristics of participants of this study. All participants are residents of the state of Pernambuco, Brazil. As for their professional positions, two work as a System Analyst, one as a Data Engineer, two as a Researcher and the other as an Information Technology Technician. We also have a Professor and a Software Engineer. All are students (masters and Ph.D.). Among the eight, three of them develop activities in the area of Data Ecosystem only in the academic environment. While the other five develop in the academic and professional environment. Regarding educational background, five have graduate degrees, one has specialization and two master's degrees. In addition, five of the eight participants have at least two years of experience in the field of data ecosystems.

The focus group was held at the Computer Center of the Federal University of Pernambuco - CIn / UFPE. The location chosen was by common agreement for all participants. The session took place in July 2019, starting at 9:00 am and ending at 11:30 am, during 2:30 hours long. To conduct the session, a moderator was chosen, who was responsible for conducting the entire session and ensuring that the focus of discussion was within the topic addressed. An observer was also part of the session, who was responsible for making all relevant notes during the session and for audio recording the discussions.

The session began with a brief presentation of the participants and mediators. Then, the discussion flowed through a predefined sequence of specific topics. The assessment itself was divided into three phases. First, the participants filled in their basic information (*i.e.* function / position, training, experience). In the second phase, the participants answered a questionnaire with twenty-four questions regarding the framework components and the model in general. This questionnaire was based on the approach used by [Luna 2009; Almeida et al. 2015; Garcia 2010] and aimed to (i) present the proposed model through brief descriptions about each component of FASED, reducing the bias that would be created from an oral presentation made by the mediators; (ii) collect quality assessment data, as well as general comments. This quality assessment data supported the discussions during the study.

6.2 Summary of Results

This section presents the analysis of the data collected in the focus group, presenting each question asked, the responses of the participants, the suggestions made by them.

We asked if each FASED indicator and its characteristics and metrics are appropriate and correctly described. The idea was to identify if the indicator and its set of characteristics and metrics are suitable for the health assessment of Data Ecosystems. In order to identify gaps, errors and/or possibilities for improvement, the participants could choose some of the following options: "no, one or more [element] need to be updated"; "No, one or more [element] need to be created"; "No, one or more [element] need to be removed". For these options, the participants were also encouraged to provide a broader detailed answer in a form of a comment.

Table IX: Focus Group General Results

Indicator	Evaluation Statements	YES	NO
Productivity	The productivity indicator is appropriate	100.00%	0.00%
	The set of productivity characteristics is appropriate	100.00%	0.00%
	The set of productivity characteristics is correctly described	62.50%	37.50%
	The set of productivity metrics is appropriate	100.00%	0.00%
	The set of productivity metrics is correctly described	87.50%	12.50%
Robustness	The robustness indicator is appropriate	100.00%	0.00%
	The set of robustness characteristics is appropriate	100.00%	0.00%
	The set of robustness characteristics is correctly described	0.00%	100.00%
	The set of robustness metrics is appropriate	62.50%	37.50%
	The set of robustness metrics is correctly described	75.00%	25.00%
Niche Creation	The niche creation indicator is appropriate	100.00%	0.00%
	The set of niche creation characteristics is appropriate	100.00%	0.00%
	The set of niche creation characteristics is correctly described	87.50%	12.50%
	The set of niche creation metrics is appropriate	100.00%	0.00%
	The set of niche creation metrics is correctly described	87.50%	12.50%
Sustainability	The sustainability indicator is appropriate	100.00%	0.00%
	The set of sustainability characteristics is appropriate	100.00%	0.00%
	The set of sustainability characteristics is correctly described	37.50%	62.50%
	The set of sustainability metrics is appropriate	87.50%	12.50%
	The set of sustainability metrics is correctly described	37.50%	62.50%

As presented in Table IX, in general, the participants were satisfied with the proposed framework and reported its importance for the context of health evaluation of DAECOs. Individually, the set of averages indicates positive results in terms of adequacy and clearness. Over 80% of responses correspond to "*Yes, the [element] is appropriate*" or "*Yes, the [element] is clearly described*". The adequacy of elements were scored with the highest values. In its turn, for most of the elements, the participants evaluation indicated some improvement to be made in terms of clearness. This seems to reflect constraints related to lack of fully knowledge on Data Ecosystem Health or even on general Data Ecosystem knowledge. Without such a background, it is crucial presenting a clear description of the characteristics and metrics. This hypothesis is grounded on the minimum set of suggestions or comments received from respondents.

The answers obtained from the first questionnaire were extremely important to identify the improvements that needed to be done in the indicators, characteristics and metrics of FASED, such as description adjustments, inclusion of new metrics and changes in the way of calculating them. For example, half of the participants suggested improving the description of the Robustness indicator, because it contained the term "survival", which was also mentioned in the description of the Sustainability indicator. This suggestion was relevant to realize that the differentiation between these two indicators was not so clear. With this, the description of the Robustness indicator was modified to

highlight its relationship with the ability to remain stable, thus resulting in the following definition "ability to remain stable when facing disruptions".

Another suggestion of a participant was to change the name of the characteristic "Management" to "Risk Management", which was accepted, since the addition of the term "risk" makes explicit the purpose of preventing. Another improvement suggestion was to modify "Centralization of Efforts" and "Decentralization of Partnerships" calculation formula. So, instead of getting answers about the existence of these aspects, they calculate the percentage of centralization of efforts and the percentage of distribution of partnerships in other countries/states/cities. Using a percentage metric provides a more relevant information regarding the balance of efforts and geographical heterogeneity of the actors, respectively.

Moreover, two participants suggested the inclusion of a new metric that would verify the existence of detailed documentation on the production and consumption processes of the ecosystem resources. This metric proposal was added to FASED, because the existence of process documentation is another way to support the execution of ecosystem activities on a daily basis and in situations of instability.

We asked to the participants to indicate to what extent the following statements hold:

—FASED is relevant for assessing the health of Data Ecosystems
—FASED is generic enough to assess the health of Data Ecosystems in any domain
—FASED is adaptable to any Data Ecosystem domain
—FASED element hierarchy is coherent

They had to indicate a value ranging from 1 to 5, where the number 1 corresponds to "Strongly Disagree" and the number 5 corresponds to "Strongly Agree". These questions were important to analyze whether the participants considered the framework hierarchy of elements to be coherent and whether the structure of the model is adaptable, generic and relevant to evaluate the health of DAECOs.

Table X: Focus Group General Evaluation

Participant	**FASED is relevant**	**FASED is generic**	**FASED is adaptable**	**FASED elements hierarchy is coherent**
Participant 1	5	5	5	5
Participant 2	5	4	4	5
Participant 3	5	5	5	5
Participant 4	5	5	5	5
Participant 5	5	4	5	5
Participant 6	4	5	5	5
Participant 7	5	5	5	5
Participant 8	5	5	5	5
Average	4.875	4.75	4.875	5

Table X shows participants' evaluation. All participants agreed that the hierarchy of FASED indicators, characteristics, attributes and metrics is coherent for evaluating the health of DAECOs, obtaining an overall average 5 on this question.

Regarding its relevance, seven of the participants rated it as 5 and one participant rated it as grade 4, resulting in an overall average of 4.875 in this question. As for its adaptability, seven of the participants rated FASED with 5, and one evaluated with grade 4, resulting in an overall average of 4,875. And finally, as for generality, five participants rated FASED with a score of 5 and two with a score of 4, resulting in an overall average of 4.75. The participants who rated the generality with grade 4 made the following comments: *"Some metrics are very specific, such as the publication of scientific papers. Depending on the domain, it does not apply. If it is to be a generic framework, all*

metrics should be described in a broad way." And another participant left the following comment: "*The adaptation of some metrics and terms may help to become more generic*".

Thus, as a result of this evaluation, it was possible to evolve the proposed model to include the most important suggestions pointed out by the participants and to identify points of improvement for future versions of FASED.

7. CONCLUSION

The framework received influences from ISO/IEC 25000:2014 and from some works identified during the ad-hoc study. This initial set of indicators, characteristics, attributes and metrics was adapted, improved, extended, refined and verified in several cycles until its presentation in the present work. Finally, this article also presents the framework evaluation results that was performed through a focus group.

As a result of the focus group, we collected evidence on the importance of the Framework for Data Ecosystems Health Evaluation, since it was possible to obtain good results and suggestions for improvements that were relevant to the framework. After this evaluation, the model was refined and some of the improvements proposed by the participants were incorporated.

As a future work, we intend to hold another focus group with other participants in order to gather new evidence. These new evidence can be used to make improvements in the framework, as for its maturity. We also intend to instantiate FASED for other data domains (*e.g.* Open Data, Private Data, Scientific Data) and apply it in these real-world scenarios where specific situations that might not have been thought of during its construction, may appear. As well, there is an intention to evolve/extend the set of metrics defined by FASED. The metrics found in the literature and those proposed by FASED are mostly quantitative. Qualitative metrics can be included to complement the evaluation of the characteristics proposed by FASED, and also to evaluate new aspects, such as the strength between relationships.

REFERENCES

Almeida, H. R., de Magalhães, E. M. C., de Moura, H. P., Teixeira Filho, J. G. d. A., Cappelli, C., and Martins, L. M. F. Evaluation of a maturity model for agile governance in ict using focus group. In *Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1*. Brazilian Computer Society, pp. 3, 2015.

Barry, M.-L., Steyn, H., and Brent, A. Determining the most important factors for sustainable energy technology selection in africa: Application of the focus group technique. *PICMET 08 - 2008 Portland International Conference on Management of Engineering Technology*, July, 2008.

Botelho, R. G. and de Oliveira, C. d. C. Literaturas branca e cinzenta: uma revisão conceitual. *Ciência da Informação* 44 (3), 2015.

Carvalho, I., Campos, F., Braga, R., David, J. M. N., Stroelle, V., and Araújo, M. A. Heal me-an architecture for health software ecosystem evaluation. *2017 IEEE/ACM Joint 5th International Workshop on Software Engineering for Systems-of-Systems and 11th Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems (JSOS)*, 2017.

Chapin III, F. S., Matson, P. A., and Vitousek, P. M. The ecosystem concept. In *Principles of Terrestrial Ecosystem Ecology*. Springer, pp. 3–22, 2011.

Chapin III, F. S., Torn, M. S., and Tateno, M. Principles of ecosystem sustainability. *American Naturalist*, 1996.

Chen, M., Mao, S., and Liu, Y. Big data: A survey. *Mobile networks and applications* 19 (2): 171–209, 2014.

Chiara, I. G. D. Grupo de foco. *Métodos qualitativos de pesquisa em Ciência da Informação*, 2005.

Chun, S., Shulman, S., Sandoval, R., and Hovy, E. Government 2.0: Making connections between citizens, data and government. *Information Polity* 15 (1, 2): 1–9, 2010.

Costanza, R. Toward an operational definition of ecosystem health. *Ecosystem health: New goals for environmental management*, 1992. *Ecosystem health: New goals for environmental management*. 239–256.

Dawes, S. S., Vidiasova, L., and Parkhimovich, O. Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly* 33 (1): 15–27, 2016.

den Hartigh, E., Tol, M., and Visscher, W. The health measurement of a business ecosystem. *European Network on Chaos and Complexity Research and Management Practice Meeting*, 2006.

Dhungana, D., Groher, I., Schludermann, E., and Biffl, S. Software ecosystems vs. natural ecosystems: learning from the ingenious mind of nature. *Proceedings of the Fourth European Conference on Software Architecture: Companion Volume*, 2010.

Easterbrook, S., Singer, J., Storey, M.-A., and Damian, D. Selecting empirical methods for software engineering research. *Guide to advanced empirical software engineering*, 2008.

Franch, X. and Carvallo, J. P. Using quality models in software package selection. *IEEE software* 20 (1): 34–41, 2003.

Franco-Bedoya, O., Ameller, D., Costal, D., and Franch, X. Queso a quality model for open source software ecosystems. *International Conference on Software Engineering and Applications*, 2014.

Franco Bedoya, Ó. H., Ameller, D., Costal Costa, D., and Franch Gutiérrez, J. Queso v2.0 a quality model for open source software ecosystems: List of measures, 2016.

Garcia, V. C. *RiSE reference model for software reuse adoption in Brazilian companies*. Ph.D. thesis, Universidade Federal de Pernambuco, 2010.

Gelhaar, J., Gross, T., and Otto, B. A taxonomy for data ecosystems. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. pp. 6113, 2021.

Group, O. G. W. Eight principles of open government data, *Open Government Working Group*, 2007. `https://public.resource.org/8\_principles.html`. Accessed in 20-April-2016.

Hanssen, G. K. and Dybå, T. Theoretical foundations of software ecosystems. *IWSECO@ ICSOB*, 2012.

Hevner, A. and Chatterjee, S. Design science research in information systems. *Design research in information systems*, 2010.

Iansiti, M. and Levien, R. Keystones and dominators: Framing the operational dynamics of business ecosystems. *The Operational Dynamics of Business Ecosystems*, 2002.

Jansen, S. Measuring the health of open source software ecosystems: Beyond the scope of project health. *Information and Software Technology* 56 (11): 1508–1519, 2014.

Kontio, J., Lehtola, L., and Bragge, J. Using the focus group method in software engineering: obtaining practitioner and user experiences. *International Symposium on Empirical Software Engineering*, 2004.

Lis, D. and Otto, B. Data governance in data ecosystems–insights from organizations, 2020.

Luna, A. J. H. d. O. *MAnGve: Um Modelo para Governança Ágil em TIC*. M.S. thesis, Universidade Federal de Pernambuco, 2009.

Manikas, K. and Hansen, K. M. Reviewing the health of software ecosystems–a conceptual framework proposal. *International Workshop on Software Ecosystems*, 2013.

Nuseibeh, B. Weaving together requirements and architectures. *Computer* 34 (3): 115–119, 2001.

Oliveira, M. I. S., Lima, G. d. F., and Lóscio, B. F. Investigations into data ecosystems: A systematic mapping study. *Knowledge and Information Systems*, 2019.

Oliveira, M. I. S. and Lóscio, B. F. What is a data ecosystem? *Digital Government Research*, 2018.

Oliveira, M. I. S., Oliveira, L. E. R., Batista, M. G. R., and Lóscio, B. F. Towards a meta-model for data ecosystems. *International Conference on Digital Government Research: Governance in the Data Age*, 2018.

Oxford Dictionary. Definition of framework, 2019.

Pollock, R. Building the (open) data ecosystem. *Open knowledge foundation Blog* vol. 31, 2011.

Silva, E., Oliveira, M., Oliveira, E., da Gama, K., and Lóscio, B. Um survey sobre plataformas de mediação de dados para internet das coisas. In *Anais do XLII Seminário Integrado de Software e Hardware*. SBC, pp. 95–106, 2015.

Soto, M. and Ciolkowski, M. The qualoss open source assessment model measuring the performance of open source communities. *International Symposium on Empirical Software Engineering and Measurement*, 2009.

Thomas, L. D. and Autio, E. The processes of ecosystem emergence. *Imperial College Business School, University of London*, 2014.

Ubaldi, B. Open government data. *OECD*, 2013.

Zaganelli, B. M., Nisenbaum, M. A., Marques, S. B., and Olinto, G. O grupo focal na ciência da informação. *Informação & Sociedade: Estudos*, 2015.

Zhang, J., Sun, J., Zhang, R., Zhang, Y., and Hu, X. Privacy-preserving social media data outsourcing. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, pp. 1106–1114, 2018.