ARTIFACT EVALUATION IN INFORMATION SYSTEMS DESIGN-SCIENCE RESEARCH – A HOLISTIC VIEW

Nicolas Prat, ESSEC Business School, Cergy-Pontoise, France, prat@essec.edu

Isabelle Comyn-Wattiau, CEDRIC - CNAM & ESSEC Business School, Paris, France, wattiau@cnam.fr

Jacky Akoka, CEDRIC - CNAM & Institut Mines Télécom - Télécom Ecole de Management, Paris, France, akoka@cnam.fr

Abstract

Design science in Information Systems (IS) research pertains to the creation of artifacts to solve real-life problems. Research on IS artifact evaluation remains at an early stage. In the design-science research literature, evaluation criteria are presented in a fragmented or incomplete manner. This paper addresses the following research questions: which criteria are proposed in the literature to evaluate IS artifacts? Which ones are actually used in published research? How can we structure these criteria? Finally, which evaluation methods emerge as generic means to assess IS artifacts? The artifact resulting from our research comprises three main components: a hierarchy of evaluation criteria for IS artifacts organized according to the dimensions of a system (goal, environment, structure, activity, and evolution), a model providing a high-level abstraction of evaluation methods, and finally, a set of generic evaluation methods which are instantiations of this model. These methods result from an inductive study of twenty-six recently published papers.

Keywords: Information Systems Research, Design Science, Artifact Evaluation, General Systems Theory, Evaluation Criterion, Generic Evaluation Method.

1 INTRODUCTION

Design-science research (DSR) supports a pragmatic research paradigm promoting the creation of artifacts to solve real-life problems (Hevner et al., 2004; Simon, 1996). Since the publication of Hevner et al. (2004), DSR has been the subject of growing attention within the IS community. It is now firmly established as a research method, while still being at the maturation stage (Fischer, 2011).

This paper focuses on artifact evaluation in DSR. Even though evaluation pervades the DSR literature, research on this topic remains at an early stage. We concur with Winter (2008) that behavioral-science research in IS significantly outperforms DSR in terms of commonly accepted, well-defined rigor standards. More specifically, *"further improvement to the criteria […] for DSR seems necessary."* (Venable, 2010). DSR in IS lacks a systematic list of evaluation criteria for artifacts and an associated set of evaluation methods. The literature presents evaluation criteria in a fragmented manner. Evaluation methods are also presented in a fragmented way and are unrelated to evaluation criteria. Consequently, DSR researchers, careful as they are about IS artifact evaluation, are often left to wonder how to perform this evaluation: what object(s) should be evaluated, according to what criteria, and what evaluation methods apply to what criteria?

Considering this research gap, the fundamental question of this paper is the *what* and *how* of IS artifact evaluation: the *what* pertains to the object(s) of evaluation and evaluation criteria, and the *how* concerns evaluation methods. We consider IS artifacts as *systems* to be evaluated. Applying general systems theory (Skyttner, 2005), we organize evaluation criteria according to the fundamental dimensions of systems, thereby providing a holistic view of artifact evaluation. The specific research questions addressed in this paper are: which criteria are proposed in the DSR literature to evaluate IS artifacts? Which ones are actually used in IS published research? How can we structure this set of criteria? Finally, which evaluation methods emerge as generic means to evaluate IS artifacts?

To conduct this research, as a complement to the literature review, we have analyzed the evaluation protocol followed in a sample of recently published IS papers applying DSR. These papers provide an indication of the commonly assessed evaluation criteria and the associated evaluation methods. Our research is itself a DSR process, resulting in an artifact including three main components. Based on general systems theory, we first propose a hierarchy of evaluation criteria for IS artifacts organized according to the fundamental dimensions of a system (goal, environment, structure, activity, and evolution). Then, we define a model aiming at a high-level abstraction of evaluation methods. Finally, we induce a set of generic evaluation methods which are instantiations of the previous model. With the hierarchy of evaluation criteria (what to evaluate), and the generic evaluation methods linked to the criteria (how to perform the evaluation), we help DSR researchers define an evaluation protocol of their IS artifact. Similarly to Venable (2010), who suggests that the Hevner-et-al guidelines should not be applied evenly to all DSR papers, we are not suggesting that all evaluation criteria should be used for all artifacts. Our approach is holistic in that it provides a systematic organization of evaluation criteria, from which DSR researchers may choose. Our analysis of papers provides an indication of the most commonly used criteria. It is also a call to find new ways (e.g. new metrics) of assessing criteria that have been deemed important in the DSR literature but are not used, or hardly ever used, in DSR practice.

The paper is organized as follows. Section 2 reviews the literature on evaluation in design-science research on the one hand and on general systems theory on the other hand. Section 3 describes our research. In Section 4, we evaluate the artifact resulting from this research. We then conclude on the contributions and limitations of this paper and sketch avenues for future work.

2 LITERATURE REVIEW

2.1 Evaluation in design-science research

Within DSR, a distinction can be made between design science (which reflects and provides guidance on artifact construction and evaluation) and design research (which constructs and evaluates specific artifacts) (Winter, 2008). Based on March and Smith (1995), Hevner et al. (2004) assert that DSR consists in building and evaluating artifacts, which may be *constructs* (concepts), *models*, *methods*, or *instantiations*. Following Walls et al. (1992), other authors, most notably Gregor and Jones (2007), argue that the ultimate goal of DSR is to produce design theories. These two views are not as antagonistic as they may seem, and may be reconciled (Gregor and Hevner, 2013) by considering design theories as a special, "upper-level" type of artifact. More specifically, design principles and requirements often constitute the central components of a design theory (Markus et al., 2002). Design principles (a.k.a. guidelines) are a sub-category of methods, and design requirements are a special type of model.

Foundational papers in the IS design-science literature stress the importance of evaluation. March and Smith (1995) provide a comprehensive list of evaluation criteria by artifact type. Hevner et al. (2004) propose some criteria and a typology of evaluation methods. The DSR process of Peffers et al. (2007) contains a demonstration and an evaluation activity. Demonstration illustrates the use of the artifact to solve one or several problem instances, and is considered as an early evaluation activity.

Some design-science papers deal specifically with evaluation. Extending the original paper by Pries-Heje et al. (2008), Venable et al. (2012) present a framework for evaluation in DSR. They characterize evaluation strategies along two dimensions: *naturalistic* versus *artificial*, and *ex ante* versus *ex post*. Ex post evaluation assesses an instantiation. Ex ante evaluation assesses an uninstantiated artifact. The framework does not consider evaluation criteria systematically, nor does it relate them to evaluation methods. Sonnenberg and vom Brocke (2012) provide multiple examples of evaluation criteria and methods, but do not relate directly evaluation methods to criteria. Cleven et al. (2009) characterize evaluation approaches along twelve dimensions, none of which mentions evaluation criteria. Peffers et al. (2012) analyze the commonly-used evaluation methods by artifact type. Järvinen (2007) argues that evaluation in DSR overemphasizes the utility criterion. Aier and Fischer (2011) present criteria for evaluating IS design theories. Siau and Rossi (2011) focus on evaluation techniques for IS analysis and design methods.

Summing up, even though the IS design-science literature acknowledges the critical role of artifact evaluation, it only provides fragmented or incomplete lists of criteria. It also presents evaluation methods in a fragmented manner, without indicating which methods apply to which criteria. This research provides a holistic view of evaluation criteria and generic evaluation methods to assess them. The holistic view is achieved by applying general systems theory.

2.2 General systems theory

In general systems theory, a system is *"an organized whole in which parts are related together, which generates emergent properties and has some purpose"* (Skyttner, 2005). The canonical form of systems (Le Moigne, 2006; Roux-Rouquié and Le Moigne, 2002) characterizes them by their five fundamental dimensions: goal, environment, structure, activity, and evolution.

Today, there is a near total consensus on which properties together constitute a theory of an open system. These twelve properties (Skyttner, 2005; von Bertalanffy, 1969) are summarized below:

- *Interrelationship and interdependence of objects and their attributes*: if elements are unrelated and independent, they cannot for a system.
- *Holism*: systems are not reducible to their parts.
- *Goal seeking*: systemic interaction must result in some goal being achieved, some final state being reached, or some equilibrium being approached.
- *Transformation process*: to achieve their goal, systems transform inputs into outputs.
- *Inputs*: in closed systems, inputs are determined a priori. On the other hand, open systems may accept new types of inputs over time.
- *Outputs.*
- *Entropy*: the amount of disorder or randomness present in the system. This property differentiates living systems from non-living systems. Living systems can reduce entropy by importing energy from their environment.

- *Regulation*: the system should adapt to the changing environment in order to achieve its goals. Regulation implies the detection and correction of deviations. Feedback is therefore a requisite of regulation.
- *Hierarchy*: systems are generally complex and decomposed into a hierarchy of sub-systems.
- *Differentiation*, i.e. division of labor.
- *Equifinality*: open systems have alternative ways of achieving the same goal from different initial states (convergence).
- *Multifinality*: from a given initial state, open systems may obtain different, and mutually exclusive, goals (divergence).

In Section 4, we will sketch how our approach addresses most of these properties.

3 FROM A HOLISTIC VIEW OF ARTIFACT EVALUATION CRITERIA TO GENERIC EVALUATION METHODS

This research is itself a design-science research. We describe our DSR process and the resulting artifact. The main components of this artifact are a holistic view of evaluation criteria (hierarchy of criteria), a model of generic evaluation methods, and examples of evaluation methods (instances of this model).

3.1 Design research process

The fundamental approach of this research is to combine DSR evaluation theory with practice. Theory was provided by the IS design-science literature. Practice was provided by a sample of design-research papers. These papers were taken from MISQ and JCIS, which are general-interest IS journals welcoming both behavioral and design-research papers. Similarly to Gregor and Hevner (2013), we selected thirteen MISQ papers. We completed our list with thirteen JCIS papers (the complete list is shown in the appendix, and the papers from the list cited in the current paper also appear in the "references" section). For MISQ as well as for JCIS, we started from the latest 2012 issue, and selected papers by proceeding backwards until we got thirteen papers. We selected only papers with a marked design-research orientation and a section on evaluation. The publication dates of the papers range from fall 2007 to 2012. Our list of MISQ papers overlaps the list of thirteen papers established by Gregor and Hevner (2013). However, the intersection is only partial because our selection includes more recent papers. More precisely, starting from their list, we chose to stick to the same number of MISQ papers but to update the list with more recently published papers. Limiting our study to a unique journal presented a bias risk. Thus, we decided to select a complementary sample in another journal markedly different from MISQ in terms of positioning and editorial policy. We chose JCIS which is also selective, but is not one of the AIS "basket of eight" journals.

Our hierarchy of criteria for DSR artifact evaluation was derived from the criteria proposed in the design-science literature. We adopted a systematic approach to build this hierarchy. To the best of our knowledge, we did not find a relevant method to obtain such a hierarchy. Therefore, we adapted the methodology for taxonomy development proposed by Nickerson et al. (2013). We first formulated the purpose of our hierarchy: our objective is to provide DSR researchers with a structured set of evaluation criteria. Then, we defined the meta-characteristic of our hierarchy that is "*the most comprehensive characteristic that will serve as the basis for the choice of characteristics*" of the hierarchy. Several difficulties in this task result from the fact that many papers propose a set of various artifacts linked together (e.g. an algorithm and an example, or a model and a method using this model, etc.). Several authors have proposed to refer to general systems theory to help structuring IS artefact variety (Matook and Brown, 2008; Gregor and Iivari, 2007). Following them, we organized the criteria holistically, along system dimensions. We elicited the criteria from the papers dealing with DSR artefact evaluation and grouped them such that criteria belong to the same group if they aim at evaluating the same facet of the system. According to Nickerson et al. (2013), this is first a conceptual-to-empirical approach and then an empirical-to-conceptual approach. The conceptual-to-empirical or top-down part consists of listing all the components of a system as the first level of the hierarchy. The empirical-to-conceptual part is a bottom-up technique taking into account each criterion and grouping it with similar ones according to the system dimension they allow us to evaluate. The process ended when no criterion was merged with a similar one or split into multiple ones in the last iteration.

We then analyzed the evaluation criteria assessed in the sample of design-research papers, to verify the correspondence between artifact evaluation theory and practice. For generic evaluation methods, we proceeded inductively: we started from the evaluation methods found in the design-research papers, abstracted them into generic evaluation methods, and then derived a model for representing generic evaluation methods.

We should acknowledge that by studying IS artifact evaluation practice from design-research papers, we don't capture the context of the papers. We base our study on the information explicitly mentioned in the papers.

The next section presents our holistic view of evaluation criteria, organized along system dimensions.

3.2 A holistic view of evaluation criteria

3.2.1 IS artifacts as systems

Simon (1996) considers design artifacts as systems. He characterizes artifacts in terms of functions (activities), goals, and adaptation (evolution). He also distinguishes the structure of the artifact from the environment in which it operates. Thus, artifacts possess all the dimensions of the canonical form of a system. IS artifacts, as specific design artifacts, can therefore be considered as systems.

In the IS design-science literature, IS artifacts also clearly appear as systems. According to Gregor (2010), *"IT artefacts are systems (or involved with systems)"*. The components of a design theory (purpose, scope, form, function, artifact mutability) (Gregor and Jones, 2007) are indeed closely related with the dimensions of systems. More specifically, *purpose* corresponds to the concept of goal, *scope* represents the frontier with the environment, *form* and *function* are respectively the structure and the activity of the artifact, and *artifact mutability* is related to evolution.

Having established that IS artifacts are systems, we compare our view of artifacts with that of March and Smith (1995). In our view, an IS artifact is the result of a DSR process. It is a system, which may be composed of sub-systems. This view is in line with Gregor and Jones (2007), whose design theory provides an integrated view of the result of a DSR process. For March and Smith (1995), a DSR process results in several artifacts, which are *constructs*, *models*, *methods*, or *instantiations*. This typology, as useful as it is, leads to a piecemeal view of design artifacts and of their evaluation.

In our system view of IS artifacts, the categories *model* and *method* from March and Smith's typology correspond to the system dimensions structure and activity (Gregor and Hevner, 2013). *Constructs* are not considered per se, but as parts of their model (structure). Finally, we concur with Gregor and Jones (2007) that *instantiations* are fundamentally different from constructs, models, and methods *("abstract artifacts")*. In particular, departing from March and Smith, we consider that instantiating abstract artifacts is a way of evaluating them, instead of viewing instantiations as artifacts to evaluate.

Summing up, IS artifacts are systems, and viewing them as such provides a holistic view of their evaluation, organizing the evaluation criteria along the fundamental dimensions of systems.

3.2.2 Hierarchy of evaluation criteria

Figure 1 shows our hierarchy of criteria. We built this hierarchy as a result of a consensus between all of us. This consensus was attained as prescribed by the hierarchical theory of justification (Laudan, 1984) based on three interrelated levels. At the factual level, we collected all criteria in the literature with their descriptions and explanations. At the methodological level, by applying linguistic rules (mainly synonym, antonym, hyponym relationships), we produced the evaluation criteria and sub-criteria. Finally, the axiological level refers, in our research, to general systems theory leading to the main system dimensions at the highest level of our hierarchy.

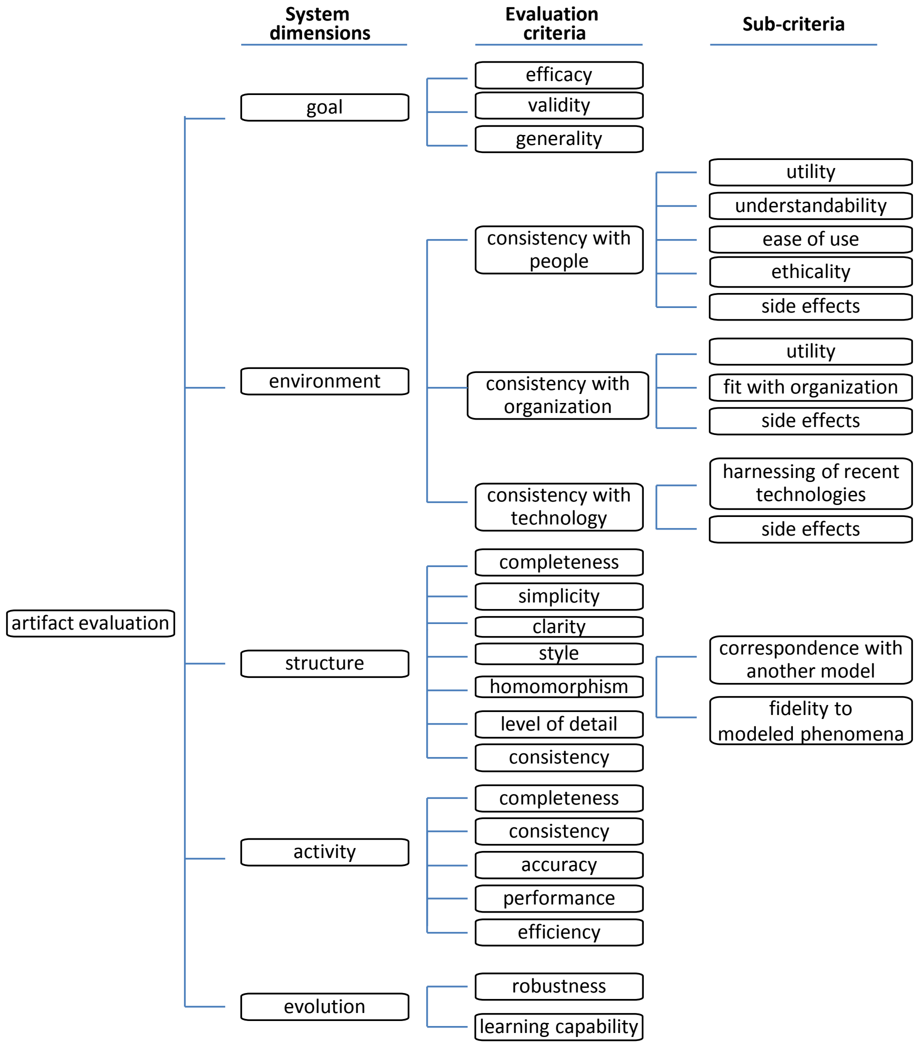

Figure 1. Hierarchy of criteria for IS artifact evaluation.

Complying with our view of IS artifacts as systems, the criteria are organized along the five dimensions of systems. In cases of synonymy, such as utility and usefulness, we selected the term most commonly used in the literature. We describe below, for each system dimension, the criteria and associated sub-criteria. Based on our literature review of DSR artifact evaluation criteria, we have defined each criterion precisely. For the sake of the present paper, we briefly summarize the criteria.

Goal is characterized by the following criteria: *efficacy* is the degree to which the artifact produces its desired effect (i.e. achieves its goal) (Venable et al., 2012). Effectiveness is sometimes distinguished from efficacy. Hevner et al. (2004) use these two terms interchangeably. We keep “efficacy”, which is the most commonly used term. Drawing on Gregor and Hevner (2013), we define *validity* as the degree to which the artifact works correctly, i.e. correctly achieves its goal. Validity encompasses reliability (Straub et al., 2004). Artifact *generality* is goal generality (Aier and Fischer, 2011): the broader the goal addressed by the artifact, the more general the artifact. Generality is often mentioned as a criterion for evaluating design theories, but is also mentioned by March and Smith (1995) for methods.

The environment of IS artifacts comprises people, organization, and technology (Hevner et al., 2004). Therefore, criteria on this dimension should verify *consistency* of the IS artifact *with people*, *organization*, and *technology*. Consistency is *“agreement or harmony of parts or features to one another or a whole”* (http://www.merriam-webster.com/dictionary). Consistency with the environment is also called external consistency (Sonnenberg and vom Brocke, 2012). The criterion *utility*, common to people and organization, measures the quality of the artifact in practical use. Utility for people does not necessarily materialize into utility for organizations. *Understandability*, *ease of use* (March and Smith, 1995), and *ethicality* (Venable et al., 2012) relate to people. *Fit with organization* (Hevner et al., 2004) characterizes the alignment of the IS artifact with its organizational environment. Within the criterion "consistency with technology", Wang and Wang (2010) argue that *"a valuable design-research artifact must be a new layer that is built on new IT artifacts"*. We call this criterion *harnessing of recent technologies*. Evaluation should also consider the *side effects* of the artifact on its environment (March and Smith, 1995).

The structure of artifacts is assessed by completeness, simplicity, clarity, style, homomorphism, level of detail, and consistency. *Completeness*, *level of detail* and *consistency* are proposed by March and Smith (1995) as criteria for models. Consistency of structure is internal consistency. March and Smith also use the criteria *simplicity* and elegance (a.k.a. *style*) for constructs. Sonnenberg and vom Brocke (2012) add the criterion of *clarity*. Even though homomorphism is not defined as a criterion in the design-science literature, it relates to construct overload, redundancy, excess and deficit (Siau and Rossi, 2011; Wand and Weber, 1993). It is also linked to the criterion of fidelity with real world phenomena (March and Smith, 1995). *Homomorphism* is the *correspondence* of a model (structure) *with another model*, or the *fidelity* of a model *to modeled phenomena*. The correspondence between a model (noted Mod1) and an ontology or, more generally, a reference model (noted Ref1), is characterized and measured by construct overload, construct redundancy, construct excess, and construct deficit. Construct overload occurs when a construct of Mod1 maps to several constructs of Ref1 (Wand and Weber, 1993). Construct redundancy occurs when several constructs of Mod1 map to the same construct of Ref1. Construct excess occurs when a construct of Mod1 does not map to any construct of Ref1. Construct deficit occurs when a construct of Ref1 does not map to any construct of Mod1.

Activity is characterized by completeness, consistency, accuracy, performance, and efficiency. *Completeness* of activity (i.e. of function) amounts to functionality (Hevner et al., 2004). *Consistency*, like completeness, applies to the dynamic aspect of artifacts (activity) as it does to their static aspect (structure). *Accuracy* and *performance* are proposed by (Hevner et al., 2004). Accuracy, as defined in (Aier and Fischer, 2011), is established when there is a demonstrated agreement with the results of existing experiments. Thus, it pertains to the dynamic aspect. Performance (e.g. speed or throughput of an activity) and *efficiency* (ratio between the outputs and inputs of the activity) are also clearly linked to this dynamic aspect.

Finally, evolution is characterized by robustness and learning capability. March and Smith (1995) propose robustness as a criterion without defining it. We characterize *robustness* as the ability to respond to the fluctuations of the environment. Thus, robustness is related to evolution. *Learning capability* is the capacity of a system to learn from its experience and the reactions of the environment.

Applying this hierarchy of criteria to our sample of papers, we obtain the following results (Figure 2).

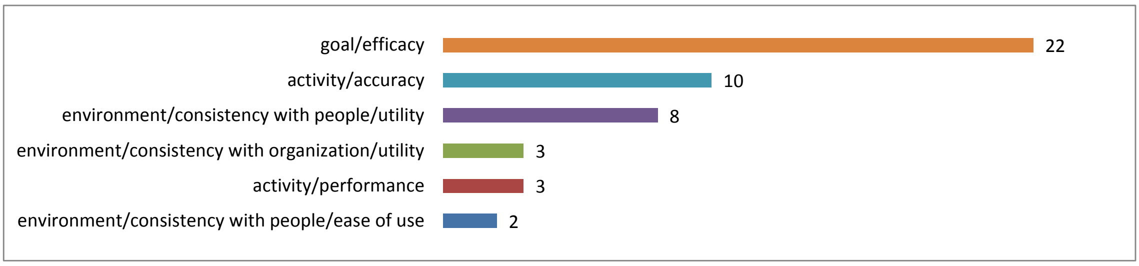

Figure 2. DSR artifact evaluation criteria used in the sample of twenty-six papers.

The first insight from this study is that the majority of criteria in our hierarchy are never evaluated in the sample: only twelve criteria are used at least once (note that Figure 2 only shows criteria that are evaluated in at least two papers). Twenty-two papers out of twenty-six evaluate the efficacy of their artifact. This result is in line with the design-science literature, which suggests demonstrating or illustrating that the artifact works in practice (i.e. achieves its goal). Eight articles evaluate the utility of the artifacts for individual users and three for organizations. The attention paid to the assessment of utility was also expected, utility being the ultimate goal of DSR. However, while utility for organizations is the ultimate measure of IS artifact relevance, utility for people is much more frequently evaluated. One possible reason is that the second type of utility is easier to assess than the first. Finally, ten papers base their validation process on accuracy (using metrics like precision, recall, or combinations of these two metrics). This corresponds to the number of papers in which the produced artifact includes an algorithm. The overall analysis of these results leads to the conclusion that the evaluation effort is concentrated on a limited set of criteria. In addition, the majority of assessed criteria are from the goal and environment dimensions. This is not surprising, considering the importance of criteria like efficacy or utility. What is more unexpected is that criteria pertaining to the structure (white-box view of artifacts, as opposed to goal and environment corresponding to a black-box view) are very rarely assessed in our sample. The majority of criteria that are never assessed belong to this dimension. This may be partly due to the choice of the two journals (IS journals, stressing the importance of efficacy and utility); a sample of computer-science journals may provide a slightly different picture. This may also reflect a lack of metrics for structure-related criteria. Finally, we should point out that the criteria of ethicality and side effects (which are, at least, partly related) are never assessed in our sample of papers. Considering the importance of these criteria in modern society (Myers and Venable, 2014; Venable, 2010), it is the role of IS as a discipline to imagine innovative approaches for assessing them.

This section has addressed the *what* of IS artifact evaluation. The object of evaluation is the artifact, considered as a system. This system can be static (i.e. include one or several models but no method) or dynamic (i.e. also include methods). The system view is opposed to the fragmented view, which considers constructs, methods, models, and instantiations as separate artifacts. This system view provides the basis for structuring evaluation criteria. In the next section, we address the *how* of artifact evaluation.

3.3 Generic evaluation methods

Regarding evaluation methods for IS artifacts, our contribution comprises a model for representing generic evaluation methods, and a set of generic evaluation methods which are instantiations of this model. The term *generic* refers to the concept of generic components in software engineering. Similarly to generic components, which are abstract and customizable for a specific application, our generic evaluation methods may have several variants depending on the specific design artifact to which they are applied. As a matter of fact, many of the evaluation methods found in our sample share common characteristics and can thus be defined as variants of the same generic evaluation method.

Our approach for defining generic evaluation methods is inductive. Starting from the evaluation methods found in the sample of design-research papers, we reduce them into generic evaluation methods sharing common characteristics. This leads us to define a model aiming at a high-level abstraction of evaluation methods. Generic evaluation methods are instantiations of this model.

For convenience of exposition, we present our model of generic evaluation methods before exhibiting examples of these methods. The model is formalized in Figure 3 as a UML class diagram.

As the model shows, a generic evaluation method assesses an evaluation criterion. Evaluation criteria are classified along system dimensions and may be decomposed into several levels, forming a hierarchy. The same criterion may be assessed by several generic evaluation methods. Generic evaluation methods vary along four fundamental characteristics, as described below: *form of evaluation*, *secondary participant*, *level of evaluation*, and *relativeness of evaluation*.

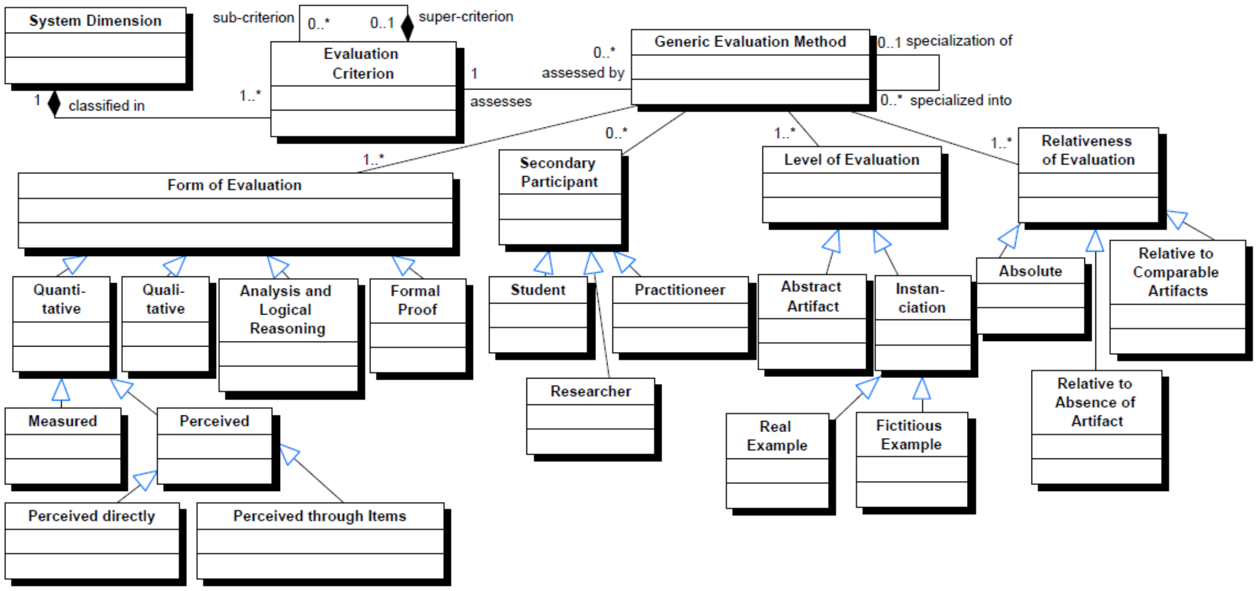

Figure 3. *A model of generic evaluation methods.*

We distinguish between *quantitative* and *qualitative* forms of evaluation (Cleven et al., 2009). Evaluation may also take the form of *analysis and logical reasoning* or *formal proof* (Hevner et al., 2004). Quantitative evaluation leads to a *measured* or *perceived* numeric value. A measure (metric) is characterized by its objectivity. A *perceived* value may be estimated *directly* or *through items*.

Secondary participants may partake in the evaluation of the IS artifact, e.g. by using a prototype and giving their feedback. They may be *students*, *practitioners*, or *researchers* (the general public may also take part in the evaluation, but this is rarely the case). Since students are often substitutes for practitioners, this typology is related to the distinction between artificial and naturalistic evaluation (Pries-Heje et al., 2008).

Evaluation may be performed at two levels: the *abstract artifact* is either assessed directly, or through one or several *instantiations*. This distinction corresponds to the distinction made by Pries-Heje et al. (2008) between ex post evaluation (instantiation) and ex ante evaluation (uninstantiated artifact). Related again to the distinction between naturalistic and artificial, these instantiations may be *fictitious* or *real examples*.

Finally, evaluation methods are characterized by the relativeness of their evaluation: the evaluation of the artifact may be *absolute* (e.g., does the artifact achieve its goal?), *relative to comparable artifacts*, or *relative to the absence of artifact* (e.g. when no comparable artifacts can be found). As claimed by Gregor and Jones (2007), IS design theories should be compared with other IS design theories with similar goals and scope. Similarly, IS artefacts should also be evaluated relatively to others artefacts.

By varying the values of the four characteristics (form of evaluation, secondary participant, level of evaluation, and relativeness of evaluation), multiple generic evaluation methods can be defined. A method can also be specialized, e.g. by specifying the value of unspecified characteristics.

Table 1 illustrates generic evaluation methods derived from the papers in our sample. Due to space constraints, the table contains only a selection of methods. The generic evaluation methods were defined inductively. For each method, some examples of papers where the method is used are shown (first column of Table 1). The methods are presented as instantiations of the model of Figure 3 and identified for further reference. When a method is the specialization of another method, the *id* of the method is followed by the *id* of its parent method (e.g. M2 *pM1*, since M1 is the parent of M2).

Examples of papers	Description of the generic evaluation method	Id	Assessed criterion	Form of evaluation	Secondary participant	Level of evaluation	Relativeness of evaluation
(Du et al., 2008; McLaren et al., 2011; Pries-Heje and Baskerville, 2008)	Demonstration of the use of the artifact with one or several examples.	M1	Goal / efficacy	Analysis and logical reasoning		Instantiation	Absolute
(Du et al., 2008; Pries-Heje and Baskerville, 2008)	Demonstration of the use of the artifact with several real examples.	M2 pM1	Goal / efficacy	Analysis and logical reasoning		Instantiation / real examples	Absolute
(Adipat et al., 2011; Wang and Wang, 2012)	Measurement of the performance of students in tasks using the artifact.	M3	Environment / consistency with people / utility	Quantitative / measured	Students	Instantiation	Absolute
(Adipat et al., 2011)	Students' perception of the utility of the artifact, based on the construct of perceived usefulness (Davis, 1989).	M4	Environment / consistency with people / utility	Quantitative / perceived / perceived through items	Students	Instantiation	Relative to absence of artifact
(Lau et al., 2012; McLaren et al., 2011)	Qualitative feedback from practitioners on the utility of the artifact.	M5	Environment / consistency with people / utility	Qualitative	Practitioners	Instantiation	Relative to absence of artifact
(Abbasi et al., 2010; Du et al., 2008; Sheikh and Conlon, 2012)	Benchmarking of the precision of an algorithm against comparable algorithms.	M7	Activity / accuracy	Quantitative / measured		Instantiation	Relative to comparable artifacts

Table 1. Some generic evaluation methods, based on the sample of twenty-six papers.

As appears in Table 1 (method M1), demonstrating the use of the IS artifact in one or several examples is a very common way of verifying if the artifact meets its goal. Method M1 corresponds to the demonstration activity in DSR processes (Peffers et al., 2007). In method M2 (specializing M1), the demonstration is based on multiple, real examples. Table 1 also exemplifies some possible methods for assessing the utility of artifacts for people. This utility may, for instance, be assessed through a performance measure, or using a quantitative or qualitative-research approach. Finally, benchmarking the precision of algorithms against comparable algorithms (method M7) is very common. Measures of recall (or other measures combining precision with recall) are also frequent.

To conclude, even if Table 1 presents only a selection of generic evaluation methods identified from the sample of papers, the study of these papers reveals great variety of such methods, in terms of form of evaluation, secondary participant, level of evaluation, and relativeness of evaluation.

4 EVALUATION OF OUR ARTIFACT

As mentioned above, this research is itself DSR. The resulting artifact comprises the hierarchy of evaluation criteria, the model for representing generic evaluation methods linked to the criteria, and examples of generic evaluation methods. Consequently, some of the evaluation criteria and methods described above may be applied to the evaluation of our artifact. Due to space limitations, this section focuses on two criteria: one related to the goal dimension (*efficacy*), and one related to the structure dimension (*homomorphism/correspondence with another model*).

4.1 Efficacy

To evaluate the efficacy of our artifact, we demonstrate its use in several real examples (i.e. we apply method M2 from Table 1).

First, the inductive stance adopted in this paper guarantees that our IS artifact has several real instantiations. More specifically, our hierarchy of criteria, by its very construction, is consistent with the criteria actually used in the sample of twenty-six papers. Similarly, the generic evaluation methods, and the model to describe them, result from abstracting evaluation methods from the sample. These evaluation methods are as many real examples illustrating the generic evaluation methods.

Second, the application of our approach to evaluate our own artifact (in terms of efficacy and homomorphism) is another application to a real example.

4.2 Homomorphism

In this section, we evaluate the correspondence with another model, a sub-criterion of homomorphism, for our IS artifact. We present generic methods (not used in the sample of papers) for assessing correspondence with another model, and apply them to our artifact.

Table 2 presents a generic evaluation method to assess construct deficit.

Description	Assessed criterion	Form of evaluation	Secondary participant	Level of evaluation	Relativeness of evaluation
Comparison of the degree of construct deficit of a model with the one of comparable models, based on the same reference model.	Structure / homomorphism / correspondence with another model	Quantitative / measured	None	Abstract artifact	Relative to comparable artifacts

Table 2. A generic evaluation method for construct deficit.

As mentioned above, the construct deficit of a model Mod1 is defined with reference to another model Ref1. It occurs when a construct of Ref1 does not map to any construct of Mod1. The degree of construct deficit is the number of constructs of Ref1 unmapped to constructs of Mod1, divided by the number of constructs in Ref1 (Recker et al., 2009). The construct deficit of Mod1 is benchmarked against the construct deficit of models comparable to Mod1, using the same reference model Ref1.

We apply the method from Table 2, and the methods for construct excess, overload and redundancy (defined similarly), to assess correspondence with another model in our approach. We assess this correspondence for our hierarchy of criteria. This hierarchy is structured along the five dimensions (five concepts) of the canonical form of a system. These dimensions, although theoretically grounded, do not map bijectively to the commonly-accepted properties of systems described in section 2.2. Therefore, we assess the correspondence of the five system dimensions (model Mod1 in the generic evaluation method of Table 2) with the commonly-accepted properties of systems (model Ref1). To this end, we map the five system dimensions with the commonly-accepted properties of systems, as represented in the left part of Figure 4. For example, the dimension Goal maps to the properties goal-seeking, equifinality, multifinality, and regulation (regulation means the “the system should adapt to the changing environment in order to achieve its *goals*”). The dimension Activity is mapped to transformation process, input and output (input and output being implicitly present in the concept of Activity). To take another example, the system property of entropy is not reflected in any of our five system dimensions. To benchmark our approach with comparable work, we consider two other IS papers that use general systems theory: one applies this theory to data quality evaluation (Orr, 1998), and the other applies it to security (Young et al., 2010). To the best of our knowledge, no other paper combines systems theory with DSR artifact evaluation, and these are the most comparable papers that we could find to our research. Similarly to our approach, we map the system dimensions used in these papers with the commonly-accepted properties of systems (Figure 4). As visible from Figure 4, the system concepts used in these approaches also map only in part with the commonly–accepted properties of systems.

Based on the mappings shown in Figure 4 for the three approaches, we can compare them regarding homomorphism with the reference model (the twelve commonly-accepted properties of systems). Let us illustrate the computation of construct deficit. For our approach, the degree of construct deficit (number of concepts from the reference model mapped to none of our five system dimensions) is 3/12=0,25. For Orr, the value is 8/12=0,75. Finally, for Young et al., the degree of construct deficit is 0,33. Thus, the system dimensions used in our work map to more concepts among the commonly-accepted properties of systems than in the other two papers, i.e. our approach is the richest in terms of representation of these commonly-accepted system properties. The price paid for this is a higher construct overload. The three approaches are equivalent in terms of construct excess and redundancy (we omit the computation details for space considerations).

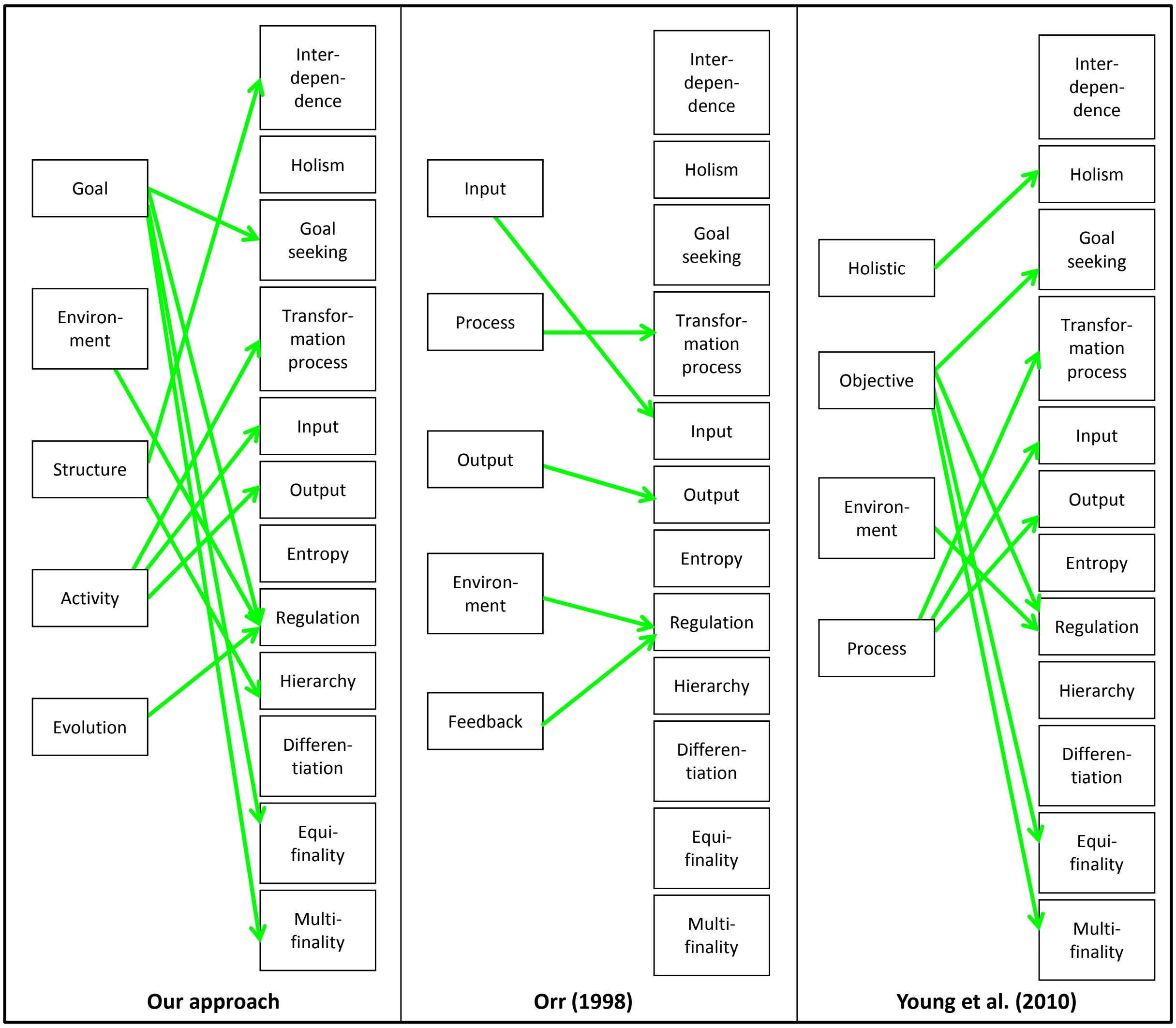

Figure 4. Mapping system dimensions used in three approaches into system properties suggested by Skyttner (2005).

5 CONCLUSION AND FUTURE WORK

Despite the key role of artifact evaluation in IS design-science research, this topic is under-explored in the literature. Consequently, DSR researchers are often left to wonder *what* to evaluate (object and criteria of evaluation) and *how* to conduct the evaluation. This paper addresses this issue by proposing a holistic view of evaluation criteria, a model providing a high-level abstraction of evaluation methods, and generic evaluation methods which are instantiations of this model.

The holistic view of evaluation criteria is obtained by considering IS artifacts as systems, and organizing criteria along the dimensions of the canonical form of a system. This holistic view, considering the IS artifact as a global system and organizing the criteria accordingly, provides an answer to the *what* of evaluation. It is a fundamental departure from the widely-accepted typology of IS artifacts, whereby the result of a DSR process is considered as a set of artifacts (constructs, models, methods, and instantiations). The characterization of IS artifacts as constructs, models, methods, or instantiations, has contributed to a piecemeal view of IS artifacts and of their evaluation. Our systems approach provides the holistic view that was missing so far in IS artifact evaluation. This holistic view is the main benefit of applying general systems theory in our research. However, we could also consider complementary theories. For example, ontologies may be relevant for unifying the vocabulary of DSR artifact evaluation. Cognitive psychology could also shed a new light on artifact evaluation (Smithson and Hirschheim, 1998).

We propose generic evaluation methods as ways of assessing the evaluation criteria. We present a model for these methods and exhibit some of them. By varying the characteristics in our model (form of evaluation, level of evaluation, etc.), new methods can be generated.

Complementary to the evaluation criteria and methods prescribed in the design-science literature, this research adopts an empirical, inductive stance. A sample of design-research papers was analyzed to confront theory with practice. This study shows that IS artifact evaluation concentrates on a small number of criteria. This calls for the development of new evaluation methods for criteria that are deemed important in the DSR literature, but seldom used in practice. Contrary to evaluation criteria, our study reveals a great variety of evaluation methods applied in DSR research. Despite this variety, these evaluation methods may be described with the fundamental characteristics proposed in our model of generic evaluation methods.

Considering again the distinction between design science and design research (Winter, 2008), our paper contribute to both. For design-science researchers, our study reveals some gaps between DSR artifact evaluation theory and practice, and suggests avenues for developing evaluation methods that are missing for some important criteria (e.g. side effects and ethics). For researchers applying designscience, the paper shows the most commonly assessed criteria, and some typical evaluation methods.

This research focuses on the evaluation of DSR artifacts, as suggested by (Hevner et al., 2004). Beyond DSR artifacts themselves (products of the DSR process), we should point out that the evaluation of the DSR process is equally important, especially in formative evaluation (Alturki et al., 2013). We keep this as an open issue for further research. Our approach could also consider in more detail the specificities of different categories of artifacts. For example, even though design theories may be considered as artifacts, their specificity may require a specialization or an extension of the criteria presented in this paper.

Despite the variety of evaluation methods found in the sample of twenty-six papers, the sample size implies that the empirical results of our study should be generalized with caution. A larger sample would give a more accurate picture of the evaluation criteria used for DSR artifacts in IS, although the general tendency is likely to be confirmed. We are currently extending the study to a larger sample of papers from a larger sample of journals. Among other things, this larger sample will enable us to investigate more deeply how criteria and evaluation methods are related, e.g. what forms of evaluation are used most frequently for what criteria. We have only started investigating the link between criteria and evaluation methods (examples are shown in Table 1). A larger sample will also enable us to further investigate the efficacy of our design artifact, by applying it to a different set of papers.

The fundamental characteristics of evaluation methods, presented in our model, could be refined by considering new characteristics or further refining existing characteristics. For example, generic evaluation methods could also be characterized by the resources that they require. Further research will refine the model of generic evaluation methods and use this model to generate new evaluation methods, particularly for the criteria which are often mentioned in the design-science literature but rarely assessed in design-research practice.

APPENDIX: SAMPLE OF DESIGN-RESEARCH PAPERS

MIS Quarterly

1. Abbasi, A., Albrecht, C., Vance, A. and Hansen, J., "Metafraud: a meta-learning framework for detecting financial fraud", 36:(4), 2012, 1293-1327.
2. Abbasi, A. and Chen, H., "CyberGate: a design framework and system for text analysis of computer-mediated communication", 32:(4), 2008, 811-837.
3. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H. and Nunamaker, J. F., "Detecting fake websites: the contribution of statistical learning theory", 34:(3), 2010, 435-461.
4. Adipat, B., Zhang, D. and Zhou, L., "The effects of tree-view based presentation adaptation on mobile web browsing", 35:(1), 2011, 99-122.
5. Adomavicius, G., Bockstedt, J. C., Gupta, A. and Kauffman, R. J., "Making sense of technology trends in the information technology landscape: a design science approach", 32:(4), 2008, 779-809.
6. Bera, P., Burton-Jones, A. and Wand, Y., "Guidelines for designing visual ontologies to support knowledge identification", 35:(4), 2011, 883-908.
7. Lau, R. Y. K., Liao, S. S. Y., Wong, K. F. and Chiu, D. K. W., "Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions", 36:(4), 2012, 1239-1268.
8. Lee, J., Wyner, G. M. and Pentland, B. T., "Process grammar as a tool for business process design", 32:(4), 2008, 757-778.
9. McLaren, T. S., Head, M. M., Yuan, Y. and Chan, Y. E., "A multilevel model for measuring fit between a firm's competitive strategies and information systems capabilities", 35:(4), 2011, 909-930.
10. Parsons, J. and Wand, Y., "Using cognitive principles to guide classification in information systems modeling", 32:(4), 2008, 839-868.
11. Pries-Heje, J. and Baskerville, R., "The design theory nexus", 32:(4), 2008, 731-755.
12. Sahoo, N., Singh, P. V. and Mukhopadhyay, T., "A hidden Markov model for collaborative filtering", 36:(4), 2012, 1329-1356.
13. VanderMeer, D., Dutta, K. and Datta, A., "A cost-based database request distribution technique for online e-commerce applications", 36:(2), 2012, 479-507.

Journal of Computer Information Systems

14. Apostolou, D., Mentzas, G. and Abecker, A., "Managing knowledge at multiple organizational levels using faceted ontologies", 49:(2), 2008, 32-49.
15. Deane, J. and Agarwal, A., "Scheduling online advertisements to maximize revenue under non-linear pricing", 53:(2), 2012, 85-92.
16. Du, H.-J., Shin, D.-H. and Lee, K.-H., "A sophisticated approach to semantic web services discovery", 48:(3), 2008, 44-60.
17. Hong, S.-Y., Kim, J.-W. and Hwang, Y.-H., "Fuzzy-semantic information management system for dispersed information", 52:(1), 2011, 96-105.
18. Hou, J.-L. and Huang, C.-H., "A model for document validation using concurrent authentication processes", 49:(2), 2008, 65-80.
19. Kim, Y. S., "Multi-objective clustering with data- and human-driven metrics", 51:(4), 2011, 64-73.

20. Li, S.-H., Huang, S.-M. and Lin, Y.-C., "Developing a continuous auditing assistance system based on information process models", 48:(1), 2007, 2-13.
21. Li, S.-T. and Chang, W.-C., "Design and evaluation of a layered thematic knowledge map system", 49:(2), 2008, 92-103.
22. Li, S.-T. and Tsai, F.-C., "Concept-guided query expansion for knowledge management with semi-automatic knowledge capturing", 49:(4), 2009, 53-65.
23. Montero, J. D., Kim, Y. S. and Johnson, J. J., "A rapid mapping conversion methodology to a commercial-off-the-shelf system", 50:(4), 2010, 57-66.
24. Sheikh, M. and Conlon, S., "A rule-based system to extract financial information", 52:(4), 2012, 10-19.
25. Ullah, A. and Lai, R., "Modeling business goal for business/IT alignment using requirements engineering", 51:(3), 2011, 21-28.
26. Wang, H. and Wang, S., "Ontology-based data summarization engine: a design methodology", 53:(1), 2012, 48-56.

References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H. and Nunamaker, J. F. (2010). Detecting fake websites: the contribution of statistical learning theory. MIS Quarterly, 34 (3), 435-461.

Adipat, B., Zhang, D. and Zhou, L. (2011). The effects of tree-view based presentation adaptation on mobile web browsing. MIS Quarterly, 35 (1), 99-122.

Aier, S. and Fischer, C. (2011). Criteria of progress for information systems design theories. Information Systems and E-Business Management, 9 (1), 133-172.

Alturki, A., Gable, G. and Bandara, W. (2013). The design science research roadmap: in progress evaluation. Proc. of PACIS 2013, Jeju Island, Korea, pp. 1-16.

Cleven, A., Gubler, P. and Hüner, K. M. (2009). Design alternatives for the evaluation of design science research artifacts. Proc. of DESRIST'09, Philadelphia, PA, pp. 1-8.

Du, H.-J., Shin, D.-H. and Lee, K.-H. (2008). A sophisticated approach to semantic web services discovery. Journal of Computer Information Systems, 48 (3), 44-60.

Fischer, C. (2011). The information systems design science research body of knowledge – a citation analysis in recent top-journal publications. Proc. of PACIS 2011, Brisbane, Australia, pp. 1-12.

Gregor, S. (2010). Building theory in a practical science. In Information systems foundations: the role of design science (Eds Gregor, S. and Hart, D.). Australian National University Press, Canberra, Australia, 51-74.

Gregor, S. and Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. MIS Quarterly, 37 (2), 337-355.

Gregor, S. and Iivari, J. (2007). Designing for mutability in information systems artifacts. In Information systems foundations: theory, representation and reality (Eds Hart, D. and Gregor, S.). Australian National University Press, Canberra, Australia, 3-24.

Gregor, S. and Jones, D. (2007). The anatomy of a design theory. Journal of the Association for Information Systems, 8 (5), 312-335.

Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28 (1), 75-105.

Järvinen, P. (2007). On reviewing of results in design research. Proc. of ECIS 2007, St. Gallen, Switzerland, pp. 1388-1397.

Lau, R. Y. K., Liao, S. S. Y., Wong, K. F. and Chiu, D. K. W. (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. MIS Quarterly, 36 (4), 1239-1268.

Laudan, L. (1984). Science and values. University of California Press.

Le Moigne, J.-L. (2006). Modeling for reasoning socio-economic behaviors. Cybernetics & Human Knowing, 13 (3-4), 9-26.

March, S. T. and Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15 (4), 251-266.

Markus, M. L., Majchrzak, A. and Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. MIS Quarterly, 26 (3), 179-212.

Matook, S., Brown, S. A. (2008). Conceptualizing the IT Artifact for MIS Research, Proc. of ICIS 2008, Paris, France, pp. 1-11.

McLaren, T. S., Head, M. M., Yuan, Y. and Chan, Y. E. (2011). A multilevel model for measuring fit between a firm's competitive strategies and information systems capabilities. MIS Quarterly, 35 (4), 909-930.

Myers, M. D. and Venable, J. (2014). A set of ethical principles for design science research in information systems. Information & Management (in press).

Nickerson, R.C., Varshney U., Muntermann J. (2013), A method for taxonomy development and its application in information systems, European Journal of Information Systems, 22 (3), 336-359.

Orr, K. (1998). Data quality and systems theory. Communications of the ACM, 41 (2), 66-71.

Peffers, K., Rothenberger, M., Tuunanen, T. and Vaezi, R. (2012). Design science research evaluation. Proc. of DESRIST 2012, Las Vegas, NV, pp. 398-410.

Peffers, K., Tuunanen, T., Rothenberger, M. A. and Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24 (3), 45-77.

Pries-Heje, J. and Baskerville, R. (2008). The design theory nexus. MIS Quarterly, 32 (4), 731-755.

Pries-Heje, J., Baskerville, R. and Venable, J. R. (2008). Strategies for design science research evaluation. Proc. of ECIS 2008, Galway, Ireland, pp. 255-266.

Recker, J., Rosemann, M., Indulska, M. and Green, P. (2009). Business process modeling – a comparative analysis. Journal of the Association for Information Systems, 10 (4), 333-363.

Roux-Rouquié, M. and Le Moigne, J.-L. (2002). The systemic paradigm and its relevance to the modelling of biological functions. Comptes Rendus Biologies, 325 (4), 419-430.

Sheikh, M. and Conlon, S. (2012). A rule-based system to extract financial information. Journal of Computer Information Systems, 52 (4), 10-19.

Siau, K. and Rossi, M. (2011). Evaluation techniques for systems analysis and design modelling methods – a review and comparative analysis. Information Systems Journal, 21 (3), 249-268.

Simon, H. (1996). The sciences of the artificial. 3rd edition. The MIT Press, Cambridge, MA.

Skyttner, L. (2005). General systems theory: problems, perspectives, practice. 2nd edition. World Scientific, Singapore.

Smithson, S. and Hirschheim, R. (1998). Analysing information systems evaluation: another look at an old problem. European Journal of Information Systems, 7(3), pp. 158-174.

Sonnenberg, C. and vom Brocke, J. (2012). Evaluations in the science of the artificial – Reconsidering the build-evaluate pattern in design science research. Proc. of DESRIST 2012, Las Vegas, NV, pp. 381-397.

Straub, D., Boudreau, M.-C. and Gefen, D. (2004). Validation guidelines for IS positivist research. Communications of the Association for Information Systems, 13 (1), 380-427.

Venable, J. (2010). Design science research post Hevner et al.: criteria, standards, guidelines, and expectations. Proc. of DESRIST 2010, St. Gallen, Switzerland, pp. 109-123.

Venable, J., Pries-Heje, J. and Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. Proc. of DESRIST 2012, Las Vegas, NV, pp. 423-438.

von Bertalanffy, L. (1969). General system theory: foundations, development, applications. Revised edition. George Braziller Inc., New-York, NY.

Walls, J. G., Widmeyer, G. R. and El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. Information Systems Research, 3 (1), 36-59.

Wand, Y. and Weber, R. (1993). On the ontological expressiveness of information systems analysis and design grammars. Information Systems Journal, 3 (4), 217-237.

Wang, H. and Wang, S. (2012). Ontology-based data summarization engine: a design methodology. Journal of Computer Information Systems, 53 (1), 48-56.

Wang, S. and Wang, H. (2010). Towards innovative design research in information systems. Journal of Computer Information Systems, 51 (1), 11-18.

Winter, R. (2008). Design science research in Europe. European Journal of Information Systems, 17 (5), 470-475.

Young, D., Conklin, W. and Dietrich, G. (2010). Re-examining the information systems security problem from a systems theory perspective. Proc. of AMCIS 2010, Lima, Peru, pp. 1-11.