Une taxonomie des méthodes d'évaluation pour les artéfacts des systèmes d'information

A Taxonomy of Evaluation Methods for Information Systems Artifacts

NICOLAS PRAT, ISABELLE COMYN-WATTIAU, AND JACKY AKOKA

Journal of Management Information Systems / 2015, Vol. 32, No. 3, pp. 229–267.

Résumé : Les artéfacts, tels que les systèmes logiciels, imprègnent les organisations et la société. Dans le domaine des Systèmes d'Information (SI), ils sont au cœur des travaux de recherche. L'évaluation des artéfacts SI représente ainsi une question essentielle. Bien que les paradigmes de recherches soient de plus en plus étroitement liés, la construction et l'évaluation des artéfacts SI ont traditionnellement été menées par la recherche dans le domaine des sciences de conception (DS, *Design Science*). La recherche en DS dans le domaine des systèmes d'information n'a pas encore atteint la maturité. C'est particulièrement vrai de l'évaluation des artéfacts. Cet article étudie le "quoi" et le "comment" de l'évaluation des artéfacts SI: quels sont les objets et les critères d'évaluation, les méthodes pour évaluer les critères, et les relations entre le "quoi" et le "comment" de l'évaluation ? Pour répondre à ces questions, nous développons une taxonomie des méthodes d'évaluation des artéfacts SI. Avec cette taxonomie, nous analysons les pratiques d'évaluation des artéfacts SI, comme ils sont reflétés par dix ans de publications dans la recherche en DS dans le panier des journaux de l'AIS (*Association for Information Systems*). Ce travail de recherche clarifie les relations importantes entre les multiples dimensions de l'évaluation des artéfacts SI, et identifie sept approches typiques d'évaluation:

- Démonstration
- Simulation, et étalonnage – basé sur des métriques – des artéfacts
- Évaluation – basée sur la pratique – de l'efficacité
- Simulation, et étalonnage – basé sur des métriques – pour l'évaluation absolue des artéfacts
- Évaluation – basée sur les pratiques – de l'utilité ou de la facilité d'usage
- Évaluation en laboratoire – basée sur des élèves – de l'utilité
- Analyse de complexité algorithmique

Cette étude révèle également un focus des pratiques d'évaluation des artéfacts SI sur un ensemble restreint de critères. Au-delà de son utilité immédiate, les chercheurs en SI sont invités à étudier les manières multiples d'évaluer l'impact organisationnel à long terme et l'impact social des artéfacts.

MOTS CLÉS : évaluation d'artéfact, analyse du contenu, évaluation de conception, science de la conception, conception de système, taxonomie.

Les artéfacts devraient être au cœur de la recherche sur les systèmes d'information (SI). Dans ce contexte, l'évaluation des artéfacts SI constitue un problème majeur. Recourir à plusieurs paradigmes de recherche, l'évaluation des artéfacts SI a traditionnellement été conduite dans la recherche en science de conception (DS, *Design Science*). La recherche en DS construit et évalue de nouveaux artéfacts. Dans le domaine des SI, la recherche en DS a commencé avec l'article de Nunamaker et al. [39], qui a proposé une approche multi méthodologique de la recherche en SI centrée sur le développement de systèmes. Même s'ils n'utilisaient pas le terme « artéfact », les auteurs ont soutenu que des modèles de données ou des méthodes par exemple, pourraient contribuer à la construction théories. La recherche en DS dans le domaine des SI a gagné une large acceptation après l'article de Hevner et al. [26]. Cet article a un rôle essentiel à jouer dans le traitement des questions organisationnelles et sociales importantes telles que la sécurité [1, 54]. Néanmoins, il faut admettre que ce paradigme de recherche mûrit encore - ses bases théoriques ne sont pas encore tout à fait stabilisées [16]. C'est particulièrement vrai de l'évaluation d'artéfacts : la recherche en DS manque de directives d'évaluation collectivement admises et spécifiques pour les différents types d'artéfacts.

Dans cet article, nous étudions le « quoi » et le « comment » de l'évaluation des artéfacts SI : Quels sont les objets de l'évaluation, et quels sont les critères par rapport auxquels les évaluer? Comment – c.-à-d. avec quelles méthodes – l'évaluation peut-elle être réalisée ? Quelle sont les relations entre les différentes dimensions dans le "quoi" et le "comment" de l'évaluation ? Pour répondre à ces questions, nous utilisons le paradigme de la recherche en DS pour construire et évaluer une taxonomie des méthodes d'évaluation pour les artéfacts SI. La taxonomie comprend six dimensions : critère, technique d'évaluation, forme d'évaluation, participants secondaires, niveau d'évaluation, et relativité de l'évaluation. Pour structurer la première dimension, nous utilisons la théorie générale des systèmes : nous considérons les critères comme des artéfacts, et organisons la hiérarchie des critères d'évaluation en fonction des caractéristiques fondamentales des systèmes. Nous appliquons la taxonomie des méthodes d'évaluation pour analyser le contenu de 121 articles de recherche en DS publiés dans le panier des huit revues de l'AIS (Association for Information Systems). La base de données qui en résulte comporte 402 méthodes d'évaluation ; elle constitue une image unique de la pratique de l'évaluation des artéfacts SI, comme en témoignent les dix années de publications dans le panier AIS. L'analyse quantitative de cette base de données améliore la compréhension de l'évaluation des artéfacts dans le domaine des SI. Cette analyse met en évidence des zones sous-explorées, où les chercheurs devraient développer de nouvelles méthodes d'évaluation en faisant appel à des paradigmes multiples (par exemple, recherche quantitative, recherche qualitative, ou recherche-action). Plus précisément, des méthodes sont nécessaires pour évaluer l'impact organisationnel à long terme des artéfacts SI et leur impact sociétal, y compris les aspects éthiques et les effets secondaires. Cette étude éclaire également les relations entre les différentes dimensions de l'évaluation et conduit à l'identification de sept approches typiques dans l'évaluation des artéfacts SI.

Littérature pertinente et questions de recherche

Nous devrions différencier la science de conception (*Design Science*, réflexion et orientation sur la construction et l'évaluation des artéfacts) de la recherche en conception (*Design Research*, construction et évaluation d'artéfacts spécifiques) [68]. Ce travail est basé sur une analyse documentaire de la science de la conception (en se concentrant sur l'évaluation des artéfacts), combinée à une analyse d'un échantillon d'articles de recherche sur la conception. Nous résumons la littérature sur les sciences de la conception et décrivons comment nous avons établi

l'échantillon d'articles de recherche sur la conception. Enfin, nous formulons nos questions de recherche pour combler les lacunes de la recherche.

Examen de la littérature en science de conception (*Design Science*)

La recherche en DS (*Design Science*, science de conception) dans les systèmes d'information (IS) construit et évalue des artéfacts, qui peuvent être des constructions, des modèles, des méthodes ou des instanciations [34]. Le premier article de SI qui relève de la recherche en DS est celui de Nunamaker et al. [39], qui propose une approche multi méthodologique de la recherche en SI, centrée sur le développement de systèmes. Hevner et al. [26] proposent un cadre et des lignes directrices pour une recherche en DS rigoureuse et pertinente. Pour compléter ces directives, la méthodologie de recherche en DS de Peffers et al. [44] spécifie le processus de recherche.

Au-delà de la construction et de l'évaluation des artéfacts SI, certains auteurs soutiennent que la recherche en DS devrait développer des théories. Le concept de théorie de la conception dans les SI a été initialement développé par Walls et al. [60]. Une théorie de la conception prescrit à la fois le produit (les propriétés qu'un artéfact devrait posséder) et le processus (la méthode de construction de l'artéfact, souvent appelée principes de conception dans d'autres publications). Les propriétés de l'artéfact doivent provenir des exigences, qui sont régies par les théories dites fondamentales (théories dans les sciences du comportement). Les hypothèses de produits testables évaluent si l'artéfact répond à ses exigences (telle que la faisabilité). Gregor et Jones [22] ajoutent deux composantes pour concevoir des théories : les construits et une instanciation d'exposition1. Le statut, la finalité et le contenu d'une théorie de la conception dans les SI restent très problématiques [56]. Dans cet article, tout en reconnaissant les théories de conception comme un résultat possible de la recherche en DS, nous nous en tenons à la typologie originale des artéfacts SI proposée par March et Smith [34]. Considérant les controverses mentionnées ci-dessus, l'ajout de théories à cette typologie est discutable. De plus, cette typologie permet déjà la représentation des composants typiques des théories de conception (par exemple, les constructions, les instanciations d'exposition, les exigences en tant que cas spécifiques de modèles, et les principes de conception en tant que cas spécifique de méthodologies).

L'évaluation est critique dans la recherche en DS. Même si le processus de recherche en DS contient un grand nombre de «micro-évaluations» [55], il est souvent divisé en activités de type construire-et-évaluer au niveau macro [26, 34, 68]. Peffers et al. [44] différencie les activités de démonstration et d'évaluation. La démonstration utilise l'artéfact pour résoudre une ou plusieurs instances de problèmes, afin d'établir que l'artéfact fonctionne. Il est suivi d'une évaluation plus formelle. L'activité de démonstration correspond à une instanciation d'exposition selon l'anatomie d'une théorie de conception de Gregor et Jones [22]. L'évaluation devrait démontrer l'utilité de l'artéfact [26]. Certains auteurs utilisent le terme « utilité perçue »[2] comme synonyme apparent de « utilité concrète »[3] [44]. Au-delà de l'utilité concrète ou perçue, plusieurs critères d'évaluation des artéfacts apparaissent dans la littérature sur les sciences de conception. L'ensemble des critères proposés par March et Smith [34] est fragmenté (organisé par type d'artéfact) et la plupart des critères ne sont pas définis. D'autres critères sont proposés [26, 50, 57], mais les listes sont incomplètes ou la signification des critères n'est pas claire. Pour évaluer les artéfacts IS, plusieurs techniques d'évaluation (par exemple, étude de cas, étude de terrain, analyse statique et simulation) sont mentionnées dans la littérature [26, 39, 43, 47]. Plusieurs articles mélangent des

[1] *expository instantiation* : une implémentation physique de l'artéfact qui contribue à illustrer la théorie, à la fois comme un support de représentation et pour des finalités de test [22, p. 322]

[2] *usefulness*

[3] *utility*

techniques d'évaluation (par exemple, des expériences de laboratoire) avec une forme d'évaluation (par exemple, des métriques) [47, 50, 57]. Certains articles corrèle le « quoi » d'une évaluation (p. ex. type d'artéfact ou critère d'évaluation) avec le « comment » de l'évaluation (p. ex. méthode d'évaluation). Venable et al. [57, 58] présente un cadre pour l'évaluation des artéfacts SI comprenant deux dimensions : naturaliste versus artificielle, et ex-ante versus ex-post. En se basant sur les « trois réalités » [52], l'évaluation naturaliste est caractérisée par des utilisateurs réels utilisant des systèmes réels (artéfacts) pour résoudre de vrais problèmes (des tâches réels dans des conditions réels). L'évaluation *ex-ante* est formative – elle a lieu pendant la construction de l'artéfact. L'évaluation *ex-post* est faite de manière cumulative – elle suit la construction de l'artéfact. Dans ce cadre, une stratégie d'évaluation est une trajectoire planifiée selon ces deux dimensions de l'évaluation. Les auteurs proposent un processus pour soutenir le choix parmi les stratégies d'évaluation. Bien que ce cadre englobe plusieurs dimensions importantes de l'évaluation dans la recherche en DS, il reste au niveau macro et ne relie pas explicitement les critères aux méthodes d'évaluation. Cleven et al. [11] proposent une taxonomie de l'évaluation des artéfacts IS. Les critères d'évaluation sont absents de la taxonomie. Peffers et al. [43] relie des types d'artéfacts avec des techniques d'évaluation. Sonnenberg et vom Brocke [50] énumèrent quelques techniques d'évaluation pertinentes pour certains critères d'évaluation, mais ne relient pas directement les critères d'évaluation aux techniques.

Cette revue de la littérature montre que DSR n'est pas tout à fait mature. Les listes de critères d'évaluation proposées dans la littérature sont fragmentées, souvent incomplètes, et la signification des critères est souvent peu claire. Les différentes dimensions des méthodes d'évaluation (le « comment » de l'évaluation) ne sont ni clairement distinguées, ni complètement spécifiées. La relation entre le « quoi » et le « comment » reste floue. Un problème particulier est le choix des méthodes d'évaluation en fonction des critères. En raison de l'absence d'une définition conceptuelle des différentes dimensions de l'évaluation, l'analyse empirique de la pratique de l'évaluation des artéfacts est difficile.

La section suivante décrit la sélection des documents de recherche en conception utilisée dans cette recherche pour analyser la pratique de l'évaluation des artéfacts SI et la composition de l'échantillon résultant.

Sélection de documents de recherche sur la conception

Pour étudier la pratique de l'évaluation des artéfacts SI, nous avons codé et analysé le contenu des documents de recherche en conception. Nous avons choisi une période de dix ans, suite à la publication de Hevner et al. [26]. Nous avons examiné les articles publiés dans le corpus de revues de l'AIS Senior Scholars: EJIS, ISJ, ISR, JIT, JMIS, JSIS, JAIS et MISQ. Ce panier a souvent été utilisé pour la citation ou l'analyse de contenu de recherche SI. Même si les revues de haute qualité en recherche en conception sont omises du panier [9], c'est une bonne source pour la recherche en DS de haute qualité en SI.

Cette étude est la première analyse approfondie et longitudinale de la pratique d'évaluation en recherche en DS basée sur le panier de revues AIS. Nous avons sélectionné tous les papiers du panier conformes aux critères décrits ci-dessous. Comparées au présent travail, les précédentes analyses de la recherche IS basées sur le panier AIS étaient muettes sur la recherche en DS [31], limitées à l'analyse de citation [16], ou centrées sur un sujet de recherche spécifique [4].

Pour sélectionner les articles, nous nous sommes limités pour aux articles et notes de recherche, à l'exclusion des éditoriaux, des articles d'opinion, des commentaires, des articles de revue et des articles publiés en ligne avant impression.

Le processus de sélection papier comprenait trois étapes: (1) vérification systématique de la table des matières de chaque revue d'avril 2004 à mars 2014, (2) recherche par mot-clé sur *Google Scholar* pour s'assurer qu'aucun papier n'avait été omis à l'étape 1, et (3) exclusion de papiers non pertinents pendant le codage du papier. L'un des auteurs a effectué les étapes (1) et (2). L'étape (3) a été réalisée collectivement (n'importe quel codeur pourrait suggérer d'exclure un papier, mais cette décision devait être approuvée par tous les codeurs de l'article).

Dans la vérification systématique des tables des matières, chaque document a été systématiquement examiné à partir du titre, et on a continué avec le résumé et le texte intégral. Dans cette première étape, un soin particulier a été pris pour ne pas exclure les articles potentiellement pertinents (si un article paraissait non pertinent après un examen plus approfondi, il pourrait être exclu dans la troisième étape). Nous avons utilisé quatre critères complémentaires, se chevauchant partiellement : trois conditions suffisantes pour l'exclusion du papier de l'échantillon et une condition suffisante pour l'inclusion du papier. La condition suffisante pour l'inclusion du papier était la mention explicite de la recherche en DS comme le paradigme de recherche principal. La première condition de l'exclusion de papier était que le principal paradigme de la recherche n'était pas la recherche en DS. Les documents de recherche qualitatifs et quantitatifs étaient souvent faciles à identifier. Nous avons également exclu les études en économie des SI et celles utilisant la recherche-action comme approche centrale. La deuxième condition de l'exclusion était que l'objectif principal du document était descriptif ou explicatif. Ce critère s'est avéré utile pour les articles qui contribuent à des modèles mathématiques, souvent sans référence explicite à leur paradigme de recherche. Si le modèle mathématique visait principalement à comprendre, nous avons exclu le papier de l'échantillon. S'il s'agissait d'un modèle pouvant être personnalisé et utilisé dans d'autres contextes (par exemple, en tant que composant d'un système d'aide à la décision), nous avons sélectionné le document. Le troisième critère d'exclusion était l'absence du papier d'un artéfact SI comme contribution centrale. Pour faciliter l'identification des artéfacts IS, nous avons appliqué notre typologie [46], qui se base sur Offermann et al. [41] et détaille les types d'artéfacts de March et Smith [34] comme suit : construit (langage, méta-modèle, concept), modèle (conception du système, ontologie, taxonomie, cadre, architecture, exigence), méthode (méthodologie, guide -ligne, algorithme, fragment de méthode, métrique), et instanciation (système implémenté, exemple).

La deuxième étape de la sélection papier consistait en une recherche par mot clé sur *Google Scholar*. Pour chaque revue, nous avons recherché n'importe où dans le document l'expression « design science » ou « design research ». Cela a donné lieu à un échantillon de 129 articles. La première étape n'avait omis que quelques documents pertinents.

Au cours de la troisième étape, nous avons excluent certains documents de l'échantillon après le début du codage des articles, si un examen détaillé révélait qu'ils étaient hors de portée de notre recherche. Lors du codage d'un article, nous avons de nouveau utilisé les deux premières conditions de l'exclusion d'articles appliquées dans la première étape. De plus, comme l'analyse de contenu d'un article a codifié sa contribution en termes d'artéfact SI et de méthodes d'évaluation, l'incapacité des codeurs à identifier au moins un type d'artéfact et une méthode d'évaluation était un signe que le document devait être exclu de l'échantillon. Après la troisième étape, la taille de l'échantillon est tombée à 121.

Le tableau 1 montre la distribution de notre échantillon par revue et par période. Il illustre une augmentation des publications de recherche sur la conception dans le panier AIS, en particulier au cours de la dernière période. Trois groupes apparaissent: JMIS (30 articles), EJIS / ISR / JAIS / MISQ

(20-22 articles) et ISJ / JIT / JSIS (2-3 articles). Le supplément en ligne A [accessible sur le site Web de l'éditeur[4]] fournit la liste complète des articles par revue.

	Apr.04–Mar. 06	Apr. 06–Mar. 08	Apr.08 – Mar. 10	Apr. 10–Mar. 12	Apr. 12–Mar. 14	Total
European Journal of Information Systems	2	8	5	2	3	20
Information Systems Journal	1	0	0	1	0	2
Information Systems Research	4	4	2	3	9	22
Journal of Information Technology	0	0	1	2	0	3
Journal of Management Information Systems	7	6	6	2	9	30
Journal of Strategic Information Systems	0	0	1	1	1	3
Journal of the Association for Information Systems	1	8	3	6	3	21
MIS Quarterly	1	0	5	6	8	20
Total	16	26	23	23	33	121

Tableau 1. Les articles de recherche en DS par revue et par période

Questions de recherche

Pour combler les lacunes de la recherche révélées par la revue de la littérature, nous répondons aux questions suivantes:

QR1 : Quels sont les objets de l'évaluation des artéfacts SI et les critères contre lequel les évaluer?

QR2 : Quelles sont les méthodes d'évaluation, c'est-à-dire les différentes options pour évaluer des artéfacts?

QR3 : Que pouvons-nous apprendre de la pratique d'évaluation des artéfacts SI par des recherches publiées? Plus précisément, quels sont les critères fréquemment évalués et les critères inexplorés? Quelles sont les relations entre les différentes dimensions de l'évaluation, par exemple, quelles méthodes d'évaluation s'appliquent à quels critères? Quelles méthodes d'évaluation typiques émergent de la pratique?

Pour répondre à ces questions, nous développons une taxonomie des méthodes d'évaluation pour les artéfacts SI, en suivant le paradigme de recherche en DS. Nous appliquons la taxonomie pour analyser le contenu de l'échantillon de documents de recherche en design.

[4] URL : http://www.tandfonline.com/doi/full/10.1080/07421222.2015.1099390

Méthode de recherche pour développer la taxonomie

Pour construire et évaluer la taxonomie (notre artéfact), nous appliquons la méthodologie pour le développement de taxonomies proposé par Nickerson et al. [38]. Nous évaluons la taxonomie formellement et de manière cumulative, en fonction des conditions d'arrêt définies par ces auteurs.

Le résultat visé : une taxonomie des méthodes d'évaluation

Une taxonomie est un ensemble de dimensions [38]. Chaque dimension consiste en un ensemble de deux ou plusieurs caractéristiques, de sorte que pour chaque objet, chaque dimension ait une et exactement une caractéristique. Cette définition simple ne permet que des dimensions plates. Nous avons également besoin de dimensions hiérarchiques, regroupant les caractéristiques (nœuds) en catégories. La catégorie la plus élevée (racine) comprend toutes les caractéristiques. Les autres catégories sont des sous-ensembles de la racine. Formellement, une taxonomie T peut être définie comme :

$$T = \left\{ Dim_i, i = 1 \ldots, n \middle| Dim_i = \left\{ Cat_{ij}, j = 1 \ldots, k_i \right\} \right|$$

$$Cat_{i1} = \left\{ Char_{im}, m = 1 \ldots, p_i; p_i \geq 2 \right\} \wedge \forall j \geq 2, Cat_{ij} \subsetneq Cat_{i1} \}$$

Par convention, la première catégorie (Cat_{i1}) est la racine. Son nom est le nom de la dimension Dim_i. Pour les dimensions plates, k_i=1. Notre taxonomie des méthodes d'évaluation comprend six dimensions:

T= {Critère; Technique d'évaluation; Forme d'évaluation, Participants secondaires ; Niveau d'évaluation, Relativité de l'évaluation}

Considérons une méthode d'évaluation et illustrons sa représentation avec cette taxonomie. Peffers et al. [44] démontrer l'efficacité de sa méthodologie de recherche en DS en l'appliquant rétroactivement à quatre projets de recherche SI déjà publiés. Dans cette méthode d'évaluation, les valeurs des six dimensions sont (dans cet ordre) : *efficacité, scénario illustratif, analyse ou raisonnement logique, aucun, exemple ou exemple réel,* et *absolu*.

Le processus de recherche

Basé sur la méthodologie proposée par Nickerson et al. [38], notre processus de recherche pour développer la taxonomie des méthodes d'évaluation est structuré comme indiqué dans la figure 1 et détaillé ci-dessous.

Choix de méta-caractéristiques, conditions d'arrêt et approche pour le développement de la taxonomie

Ce travail étudie le «quoi» et le «comment» de l'évaluation, et la façon dont ils sont corréler. Par conséquent, pour la taxonomie des méthodes d'évaluation des artéfacts SI, le «quoi» et le «comment» sont les *méta-caractéristiques*.

Il existe trois types de conditions d'arrêt [38] : les conditions d'arrêt qui font partie de la définition d'une taxonomie, les conditions d'arrêt objectives et les conditions d'arrêt subjectives. Dans notre cas, les conditions d'arrêt applicables dépendent de la phase de développement de la taxonomie. Le tableau 2 montre les conditions d'arrêt applicables dans notre approche, et les phases du processus de développement de la taxonomie (tel que représenté dans la Figure 1) où ces conditions s'appliquent. Des détails sur l'applicabilité des conditions d'arrêt dans les différentes phases sont fournis ci-dessous.

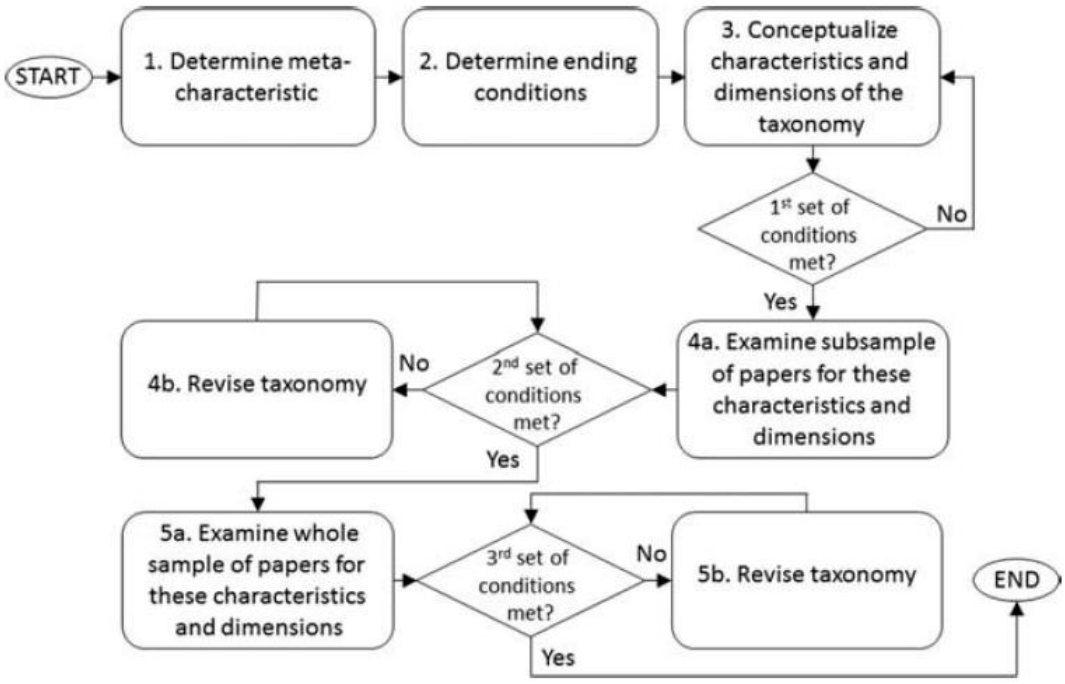

Figure 1 : Méthodologie pour développer la taxonomie des méthodes d'évaluation pour les artéfacts SI

Il existe deux approches possibles pour le développement d'une taxonomie: du conceptuel vers l'empirique, et de l'empirique vers le conceptuel. Dans cet article notre approche est du conceptuel vers l'empirique : nous conceptualisons une version initiale de la taxonomie (phase 3), et l'utilisons pour analyser et comprendre la pratique de l'évaluation des artéfacts à travers l'examen des documents de recherche en design. Cet examen est basé sur l'analyse de contenu [64]. Avant d'examiner notre échantillon complet de documents (phase 5), nous examinons un sous-échantillon pour tester le schéma de codage (phase 4). À chaque phase du développement de la taxonomie, nous testons les conditions applicables et révisons la taxonomie en conséquence.

Conceptualisation des caractéristiques et des dimensions de la taxonomie

Cette phase utilise la littérature scientifique sur l'évaluation des artéfacts pour conceptualiser la taxonomie. Dans cette phase de conceptualisation, la plupart des conditions de fin du tableau 2 s'appliquent. Les conditions de fin objectives liées à l'examen ou à la classification des objets ne sont pas applicables. Nous considérerons les conditions d'arrêt subjectives comme «robustes» et «explicatives» une fois la hiérarchie finalisée (évaluation de manière cumulative).

Examen d'un sous-échantillon de documents et révision de la taxonomie

Comme recommandé par la littérature sur l'analyse de contenu, cette phase teste le schéma de codage sur un sous-échantillon d'articles, ce qui entraîne des révisions de la taxonomie et des heuristiques de codage (voir la section «Méthode d'analyse de contenu»).

Examen de l'ensemble des documents et révision de la taxonomie

Cette phase analyse le contenu de l'ensemble de l'échantillon, ce qui donne une base de données Microsoft Access de **402 méthodes d'évaluation**. Cette base de données est ensuite analysée quantitativement pour comprendre la pratique de l'évaluation des artéfacts SI, avec des analyses statistiques avancées la plupart effectuées avec R [45]. Les résultats de cette phase fournissent également une base pour affiner davantage la taxonomie (plus précisément, la hiérarchie des critères). Après cette phase, nous évaluons la taxonomie sommairement.

Les sections suivantes détaillent les phases de construction et d'application de la taxonomie (phases 3, 4 et 5 de la figure 1). Nous examinons les conditions d'arrêt applicables dans chaque phase, les conditions de fin «subjectives» [38] étant en effet des critères d'évaluation, utilisés pour l'évaluation formative et de manière cumulative de la taxonomie.

Conceptualisation des caractéristiques et des dimensions de la Taxonomie

Cette phase (phase 3 de la figure 1) utilise la littérature scientifique sur la conception des SI pour l'évaluation des artéfacts. Sur la base de la méta-caractéristique, la taxonomie résultant de cette conceptualisation comporte six dimensions:

T = {Critère; Technique d'évaluation; Forme d'évaluation, Participants secondaires ; Niveau d'évaluation, Relativité de l'évaluation}

La première dimension concerne le «quoi» et les autres détaillent le «comment». La méthode d'évaluation est une combinaison unique de caractéristiques pour les six dimensions. Dans les sections suivantes, nous discutons du «quoi» et du «comment» de l'évaluation, et la vérification des conditions de fin pour la taxonomie conceptualisée.

Le "quoi" de l'évaluation

Le "quoi" concerne les objets de l'évaluation et les critères d'évaluation de ces objets. Nous soutenons que les sorties de la recherche en DS sont des systèmes d'artéfacts SI. Cette vision systémique constitue la base de l'organisation des critères d'évaluation.

Les résultats de la recherche en DS en tant que systèmes d'artéfacts

Simon [48] décrit les artéfacts en termes de leur fonctionnement et organisation. Il considère les artéfacts complexes comme des hiérarchies et souligne leurs opérations internes et leurs interactions avec l'environnement. Même s'il pointe certaines limites de la théorie des systèmes généraux (*General System Theory*, GST), sa description des artéfacts complexes reflète une vision systémique. Si les artéfacts complexes sont des systèmes, les artéfacts SI ne doivent pas faire exception. Cette vision est également soutenue dans la littérature SI. Par exemple, Gregor et Iivari [21, p. 6] affirment que « *les SI et les artéfacts informatiques sont tous les deux considérés comme un système parce qu'ils ont quelque part dans leurs limites un système informatique qui permet à l'artéfact de changer et d'afficher la mutabilité*. »

La vue des systèmes sous-tend également les composants des théories de conception des SI [22], en particulier le but et la portée, les principes de forme et de fonction, et la mutabilité artéfact.

Certes, certains artéfacts SI peuvent ne pas constituer des systèmes en eux-mêmes. Cependant, en fin de compte, un effort de recherche ne DS devrait produire au moins une instanciation et un artéfact abstrait (construit, modèle ou méthode), formant un système d'artéfacts. Parmi les typologies proposées dans la littérature sur la GST, deux sont pertinentes ici [2,49] : la distinction entre les systèmes concrets et abstraits, et entre les systèmes dynamiques et statiques. Les systèmes concrets ont une existence physique. Les systèmes d'abstraction sont faits de concepts. En ce sens, les théories peuvent être considérées comme des exemples de systèmes abstraits. Les éléments clés d'une théorie sont les concepts et les énoncés des relations entre eux [19]. Les systèmes dynamiques possèdent des composants structurels et des activités. Les systèmes statiques n'effectuent en eux-mêmes aucune activité. La GST se concentre souvent sur des

systèmes concrets et dynamiques (par exemple, des systèmes de vie). Néanmoins, la vue des systèmes reste pertinente pour les autres catégories. Par exemple, les systèmes abstraits peuvent avoir un environnement et subir une évolution, même si l'évolution et l'environnement ont des sémantiques différentes pour cette catégorie de systèmes [2,49].

Condition d'arrêt	**Phases**
Définition d'une taxonomie	
La taxonomie est constituée de dimensions chacune avec des Caractéristiques	3,4,5
La taxonomie est constituée de dimensions ayant chacune une Caractéristiques	3,4,5
Conditions d'arrêt objectives	
Tous les objets ou un échantillon représentatif d'objets ont été examinés	5
Au moins un objet est classé sous chaque caractéristique de chaque dimension	5
Chaque dimension est unique et non répétée (c'est-à-dire qu'il n'y a pas de dimension reproduction)	3
Chaque caractéristique est unique dans sa dimension (c'est-à-dire qu'il n'y a pas de duplication dans une dimension)	3,4,5
Conditions d'arrêt subjectives	
Concis: nombre limité de dimensions et nombre limité de caractéristiques	3,5
Robuste: assez de dimensions et de caractéristiques pour différencier les objets	5
Complet: toutes les dimensions et caractéristiques nécessaires pour classer les objets d'intérêt	3,5
Extensible: inclusion facile de dimensions et de caractéristiques supplémentaires	3,4,5
Explicatif: explications utiles de la nature des objets	5

Tableau 2. Conditions d'arrêt appliquées dans le développement de la taxonomie

Les typologies des systèmes mentionnés ci-dessus sont liées à la typologie des artéfacts [34]. Les instanciations sont des systèmes concrets. Construits, modèles et méthodes sont des systèmes abstraits: ils « *n'ont pas d'existence physique, sauf qu'ils doivent être communiqués en mots, en images, en diagrammes ou en d'autres moyens de représentation* » [22, p. 321]. Les constructions et les modèles sont des systèmes statiques. Les méthodes sont dynamiques.

En considérant les sorties des recherches en DS comme des systèmes d'artéfacts SI, nous devrions évaluer ces sorties en fonction des propriétés fondamentales des systèmes. Un système est un ensemble de composants en interrelation entre eux et avec leur environnement [59,65]. La GST explique et décrit le fonctionnement et l'évolution des systèmes en tant que tels. Elle considère les systèmes comme holistiques, ouverts, axés sur les objectifs et auto-organisés. Un système complexe est constitué de sous-systèmes organisés hiérarchiquement. Les sous-systèmes communiquent et travaillent ensemble pour s'adapter à l'environnement. Plusieurs théories systémiques ont été proposées [49]. De plus, Bowler [7] caractérise un système comme une hiérarchie de sous-systèmes partageant certaines caractéristiques communes. De même, Churchman [10] met l'accent sur le rôle des utilisateurs, des décideurs et des concepteurs de systèmes. Toutes ces approches partagent les douze propriétés fondamentales des systèmes [49] : interrelation et interdépendance des objets, holisme, recherche d'objectifs, processus de transformation, entrées, sorties, entropie, régulation, hiérarchie, différenciation, équifinalité et multi finalité. Afin de réduire ces propriétés fondamentales à un ensemble plus facile à gérer, nous les mappons dans les cinq aspects définissant la forme canonique d'un système [30], respectivement: objectif, environnement, structure, activité et évolution. Ces cinq aspects englobent toutes les caractéristiques d'un système sur lequel il existe un consensus parmi les

auteurs mentionnés ci-dessus. Le tableau 3 illustre la cartographie entre les propriétés fondamentales des systèmes et les cinq aspects constituant la forme canonique d'un système (on omet l'entropie car cette propriété n'applique que des systèmes tolérants). Chacune des onze propriétés fondamentales des systèmes correspond à au moins un aspect de la forme canonique d'un système. Ainsi, cette forme canonique suffit à structurer l'évaluation des sorties des recherches en DS. Elle servira de base pour l'organisation des critères d'évaluation.

En effet, nous considérons les objets de l'évaluation dans la recherche en DS comme des systèmes d'artéfacts SI. Ces artéfacts peuvent être concrets ou abstraits et dynamiques ou statiques. La visualisation des sorties de la recherche en DS en tant que systèmes permet une vue holistique de leur évaluation : même si la recherche en DS produit plusieurs artéfacts différents, nous ne devrions pas considérer ces artéfacts isolements, et les méthodes d'évaluation évaluent souvent le système dans son ensemble plutôt qu'un artéfact spécifique. La vue des systèmes fournit également la base pour organiser la hiérarchie des critères, selon les cinq aspects de la forme canonique d'un système (objectif, environnement, structure, activité et évolution). En fonction de la nature des artéfacts composant l'évaluateur dans la recherche en DS (concret versus abstrait et dynamique versus statique), certains critères peuvent ne pas être pertinents. Par exemple, les critères de l'aspect de la réactivité ne s'appliquent qu'aux systèmes dynamiques. Néanmoins, considérer DSR évalue que les systèmes fournissent une perspective perspicace sur leur évaluation. Structurer la hiérarchie des critères selon les cinq aspects des systèmes ne signifie pas que nous devrions évaluer les DSR selon tous les critères de tous les aspects. L'analyse des documents de recherche sur la conception (section «Résultats de l'analyse des documents de recherche sur la conception») révélera les critères les plus fréquemment évalués et examinera les relations entre ces critères et d'autres dimensions de l'évaluation.

	Objectif	Environnement	Structure	Activité	Évolution
Interrelations et interdépendance			X		
Holisme			X		
Recherche de l'objectif	X				
Processus de transformation				X	
Entrées		X		X	
Sorties		X		X	
Règlement		X			X
Hiérarchie			X		
Différenciation				X	
Equifinalité	X				
Multifinalité	X				
Sources: Pour les propriétés fondamentales, voir [49]; pour la forme canonique, voir [30].					

Tableau 3. Propriétés fondamentales par rapport à la forme canonique des systèmes

Hiérarchie de critère d'évaluation

Nous appliquons le triage de carte[5] pour construire la hiérarchie de critères. Les bénéfices de cette méthodologie incluent la subjectivité réduite. Elle est relativement simple, permet aux chercheurs de mieux comprendre l'organisation de l'information. Elle assume que les participants aient des connaissances dans le domaine d'information en question. La littérature décrit des méthodes qualitatives d'analyse des donnée de triage de carte, spécialement lorsque le nombre de participants est léger. Initialement, nous construisons une liste des articles les plus cités dans la recherche en DS. Nous les classons par ordre décroissant des citations de *Google Scholar*. Nous extractions 71 critères d'évaluation de ces pages. Nous créons une carte (document Word) pour chaque critère mentionné dans chaque page. Il contient a nom (le nom indiquant le critère dans l'article), une source (la référence de l'article), une définition (si fournie), et un commentaire (n'importe quelle information supplémentaire du critère issus de l'article).

Les trois auteurs ont réalisé le triage de carte. Les théories des systèmes nous a permis d'esquisser les catégories pour structurer l'ensemble des critères d'évaluation. Ainsi, en accordance avec les cinq aspects des systèmes identifiées auparavant, un triage fermée permettant la catégorisation des 71 critères dans cinq ensemble: objectifs, environnement, structure, activité et l'évolution. La catégorie « environnement » est divisée en trois sous catégories : humaines, organisation, et la technologie [26]. Pour la première itération, tous les chercheurs ont d'abord effectué un tri individuel. Ils ont convenu d'un ensemble d'instructions standard : ils devaient trier chaque carte en une ou plusieurs catégories. Il n'y a pas de placement correct prédéfini. Par conséquent, nous ne pouvions pas utiliser un taux de réussite pour mesurer la qualité de l'accord entre les codeurs. Nous avons défini une mesure comme suit : si les participants n'étaient pas d'accord sur la catégorisation d'une carte, la mesure était nulle ; sinon, la mesure était n si n participants étaient d'accord sur la catégorie. Enfin, nous avons comparé cette mesure à l'accord total des trois codeurs sur les 71 critères, conduisant à un ratio de 80%. Aucune autorité générale n'existe en ce qui concerne les scores requis [62]. Pour améliorer la qualité de nos résultats, nous avons mené une séance de brainstorming afin de confronter les désaccords et de trouver la meilleure catégorisation. La deuxième itération visait à réduire le nombre de cartes. En raison de nombreux noms identiques, nous avons dû vérifier si des critères de noms identiques ou synonymes pouvaient être fusionnés (pour respecter la condition de fin que chaque caractéristique d'une taxonomie soit unique dans sa dimension) ; nous devions également détecter les homonymes potentiels. Nous avons effectué une comparaison individuelle des critères situés dans les mêmes catégories grâce à la synthèse de la première itération. Sur la base des relations linguistiques, nous avons proposé une fusion lorsque, en lisant les définitions, deux ou plusieurs critères semblaient être des synonymes ou des antonymes (par exemple, la simplicité et la complexité). Enfin, la troisième itération a choisi le meilleur nom pour les synonymes et le terme positif dans le cas des antonymes (par exemple, la simplicité par rapport à la complexité). Si un terme était l'hypéronyme d'un autre, nous avons enrichi la hiérarchie avec ce lien. Nous fournissons la hiérarchie finale dans l'annexe, dans le cadre du système de codage utilisé dans l'analyse du contenu de l'échantillon de documents. L'annexe présente les définitions de tous les critères, en citant les documents à partir desquels nous avons pris ou adapté les définitions. La hiérarchie des critères d'évaluation constitue la première dimension de notre taxonomie. La spécification de cette dimension suit (les caractéristiques sont en italique) :

Critère = {But (atteinte de l'objectif (*efficacité, efficacité, validité*), utilité, faisabilité (...), généralité) ; Environnement (...) ; Structure (...) ; Activité (...) ; Évolution (…)}

[5] Triage de carte : *card sorting*

Le "comment" de l'évaluation

La littérature de recherche en DS distingue l'évaluation *naturaliste* de l'évaluation *artificielle* [57,58]. Dans le premier cas, les vrais utilisateurs utilisent de vrais artéfacts pour résoudre de vrais problèmes (de vraies tâches dans des contextes réels) [52]. Plusieurs dimensions opérationnalisant le «comment» de l'évaluation reflètent cette distinction entre évaluation naturaliste et artificielle.

Technique d'évaluation

La technique d'évaluation est une dimension fondamentale [11, 26, 39, 47]. Nous la définissons comme une dimension hiérarchique, en adaptant légèrement la typologie de Hevner et al. [26]. Notre catégorie «basée sur des questions» comprend l'enquête [47] et le groupe de discussion [53]. Nous ajoutons également la recherche-action. Considérant que les tests sont marginaux parmi les méthodes d'évaluation [4], nous ne les décomposons pas en tests en boîte noire ou en boîte blanche.

Technique d'évaluation = {observationnelle ou participative (*étude de cas, étude de terrain, recherche-action*) ; Analytique (*analyse statique, analyse dynamique*) ; Expérimental (expérience contrôlée (*expérience de laboratoire, expérience sur le terrain*), simulation) ; Essai ; Descriptif (*argument éclairé, scénario illustratif*) ; Basé sur des questions (*Enquête, groupe de discussion*)}

Forme d'évaluation

Comme mentionné précédemment, plusieurs articles scientifiques de conception mélangent la technique d'évaluation (par exemple, une expérience de laboratoire) et la forme d'évaluation (par exemple, la métrique). Pour se conformer à l'exigence que les caractéristiques de chaque dimension d'une taxonomie soient mutuellement exclusives [38] (par exemple, pour exprimer qu'une métrique est utilisée dans le contexte d'une expérience de laboratoire), la technique et la forme d'évaluation doivent être distinguées.

Cleven et al. [11] distinguent les approches quantitatives et qualitatives de l'évaluation. Nous divisons les variables quantitatives en variables mesurées et perçues (alias latentes). Les métriques sont cruciales dans l'évaluation [34]. Les variables perçues peuvent être perçues directement ou à travers des items (constructions latente-formative ou latente-sommative). D'autres formes d'évaluation sont l'analyse [26] et le raisonnement logique [50], et la preuve formelle [26].

Forme d'évaluation = {quantitatif (*mesuré*, perçu (*perçu directement, perçu à travers les items*)) ; *Qualitatif ; Analyse et raisonnement logique, Preuve formelle*}.

Participants secondaires

Les participants secondaires participent à l'évaluation des artéfacts sans être directement impliqués dans leur construction. Ils peuvent être des étudiants, des praticiens ou d'autres chercheurs. Cette dimension concerne la distinction entre évaluation *naturaliste* et évaluation *artificielle* (vrais versus utilisateurs faux) Les étudiants fournissent souvent des conditions moins réalistes d'évaluation que les praticiens [66]. Une méthode d'évaluation peut impliquer plusieurs types des participants secondaires. Par conséquent, nous considérons toutes les combinaisons possibles satisfaire à l'exigence selon laquelle les caractéristiques d'une dimension doivent être mutuellement exclusif.

Participants secondaires = {*Étudiants ; Praticiens ; Chercheurs ; Étudiants et praticiens ; Chercheurs et praticiens ; Étudiants et chercheurs ; Élèves ; Chercheurs et praticiens ; Aucun*}

Niveau d'évaluation

Cette dimension correspond à la distinction entre évaluation ex-ante (évaluation d'un artéfact abstrait) et évaluation ex-post (évaluation d'un artéfact) [57]. Une instanciation peut être une application de l'artéfact à un "réel problème " [52] ou à un fictif. Nous utilisons le terme "exemple" de préférence pour "Problème" car un artéfact évalué (par exemple, un algorithme) peut être instancié sur un ensemble de données sans référence explicite à un problème.

Niveau d'évaluation = {*Artéfact abstrait*; Instanciation (*Exemple ou exemples fictif(s); Exemple ou exemples réel(s))}*

Relativité de l'évaluation

Un nouvel artéfact devrait être meilleur que les artéfacts existants [39]. L'évaluation devrait donc établir sa supériorité en comparant la solution à celles existantes. La performance réalisée avec un nouvel artéfact peut également être comparée à la performance sans artéfact.

Relativité de l'évaluation = {*Absolue ; Relative à l'absence d'artéfact ; Relative à artéfacts comparables*}

Conditions d'arrêt et évaluation formative de la taxonomie

La conceptualisation des caractéristiques et des dimensions de la taxonomie a été itératif. C'était particulièrement le cas pour la hiérarchie des critères d'évaluation. En définissant les dimensions, nous avons systématiquement vérifié que les caractéristiques de ces dimensions étaient exclusives. Pour chaque dimension, nous avons également surveillé les caractéristiques manquantes pour garantir l'exhaustivité. De plus, par souci d'exhaustivité, la conceptualisation de la taxonomie a été basée sur tous les documents scientifiques de conception pertinents traitant de l'évaluation d'artéfact SI. A la fin de la conceptualisation des caractéristiques et dimensions de la taxonomie, cette taxonomie semble comprendre des dimensions avec mutuellement exclusives et des caractéristiques collectivement exhaustives. Cela nécessitera une confirmation dans les phases empiriques de développement de la taxonomie (examen des documents de recherche en conception). Nous devrions soulignons également que la notion d'exhaustivité est relative. Nous avons analysé la littérature scientifique pertinente en science de conception sur l'évaluation des artéfacts, mais on ne peut pas définitivement affirmer l'exhaustivité de cette littérature en ce qui concerne les six dimensions de la taxonomie.

Les six dimensions de la taxonomie sont uniques. Cela n'implique pas qu'ils sont orthogonaux. Par exemple, méthodes d'évaluation expérimentales (technique d'évaluation) opère sur des instanciations (niveau d'évaluation).

Dans la phase de conceptualisation, trois conditions d'arrêt subjectives s'appliquent : la taxonomie devrait être concise, complète et extensible. En termes de notre hiérarchie de critères d'évaluation, la taxonomie devrait être simple (structure / simplicité), complète (structure / exhaustivité), et modifiable (évolution / modification). La modifiabilité est hypéronyme d'extensibilité. Nous évaluons formellement ces trois critères ci-dessous

Pour évaluer la simplicité, nous proposons une métrique spécifique (voir supplément B en ligne). Selon cette métrique, la simplicité de la taxonomie résultant de la la phase de conceptualisation est de 0,21. La hiérarchie des critères d'évaluation est la plus complexe dimension de la taxonomie. La simplicité de cette dimension, prise individuellement, est 0,26. Nous calculerons à nouveau ces mesures à la fin du développement de la taxonomie processus (évaluation de manière cumulative).

A ce stade, la complétude de la taxonomie des méthodes d'évaluation est assurée par l'examen systématique de la littérature sur la science de la conception des SI sur l'évaluation des artéfacts.

Ce n'est qu'un aspect de la complétude. Nous revérifierons ce critère dans la phase empirique du développement de la taxonomie.

Enfin, la structure hiérarchique de la taxonomie facilement modifiable en permettant d'ajouter, de supprimer ou de fusionner des éléments à un niveau donné de cette organisation hiérarchique.

Méthode d'analyse de contenu et test initial du schéma de codage

Dans la méthodologie de développement de la taxonomie des méthodes d'évaluation, les phases empiriques (phases 4 et 5 de la figure 1) appliquent la taxonomie à l'analyse du contenu de documents de recherche en conception. La phase 4 teste le schéma de codage sur un sous-échantillon d'articles, et révise la taxonomie et les heuristiques de codage en fonction des résultats de ce test. Dans cette section, nous détaillons la phase 4 dans le contexte de la méthode suivie d'analyse de contenu. La phase 5 sera décrite dans les sections suivantes.

Méthode analyse de contenu

Protocole de codage

L'analyse de contenu comble le fossé entre les données qualitatives et les analyses quantitatives [12]. Il utilise un schéma de codage avec des unités de codage, des catégories et des règles de codage. Le codage attribue une catégorie à une unité. Les règles de codage spécifient comment procéder. Le développement d'un schéma de codage nécessite plusieurs étapes [64] : (1) définir les unités d'enregistrement, (2) spécifier les catégories et les règles de codage, (3) tester le codage sur un sous-échantillon de texte, (4) évaluer la fiabilité, (5) si la fiabilité est faible, réviser es règles et les catégories de codage, et passer à l'étape 3 si nécessaire, (6) coder l'ensemble de l'échantillon, et (7) évaluer la fiabilité.

Dans cette recherche, nous avons deux niveaux d'unités de codage. Le premier niveau est l'article de journal. Pour chaque article de notre échantillon, nous avons codé l'artéfact ou les artéfacts SI fournis par l'article, en utilisant la typologie des artéfacts [46]. Le deuxième niveau est la méthode d'évaluation. Pour chaque article, nous avons codé chaque méthode utilisée par les auteurs pour évaluer les artéfacts, en fonction de notre taxonomie des méthodes d'évaluation des artéfacts. Ainsi, la taxonomie était au centre du schéma de codage. Le codage d'une méthode d'évaluation a déterminé la caractéristique («catégorie», dans le vocabulaire de l'analyse du contenu) pour les six dimensions de la méthode d'évaluation. L'annexe montre un extrait du schéma de codage. Pour garantir la fiabilité, nous avons également défini des heuristiques et des exemples de codage détaillés. Ce travail a nécessité une série d'ateliers avec tous les codeurs et a donné lieu à un document PowerPoint de soixante et onze diapositives.

Nos unités d'enregistrement ont exclu l'utilisation de logiciels d'analyse de contenu. Plus spécifiquement, les termes désignant les critères d'évaluation (par exemple, "performance") sont souvent ambigus, ce qui nécessite de revenir constamment aux définitions des critères dans le schéma de codage. Les trois auteurs ainsi qu'un assistant de recherche ont participé au codage. Le codage a été effectué avec Microsoft Access. Les bases de données Access individuelles ont ensuite été utilisées pour le calcul automatisé de la fiabilité de l'intercodeur, et la base de données intégrée pour l'analyse des données a été dérivée de ces bases de données une fois que les différences d'intercodeur ont été discutées et résolues.

Nous avons testé le schéma de codage sur un sous-échantillon de dix articles. Le sous-échantillonnage a pris en compte les proportions d'articles dans les différentes revues et a cherché la diversité dans les artéfacts apportés et les méthodes d'évaluation. Chaque auteur et l'assistant

de recherche ont codé les dix documents, et la fiabilité de l'intercodeur a été calculée. Sur la base des résultats du test, la taxonomie et les heuristiques de codage ont été révisées. Le schéma de codage modifié a ensuite été appliqué à l'échantillon entier. Dans cette phase, tous les articles étaient codés en double. Pour assurer la continuité dans le codage et dans les discussions sur les désaccords, un auteur a codé tous les articles. Après double codage, la fiabilité a été calculée à nouveau. Nous détaillons l'évaluation de la fiabilité et l'amélioration ci-dessous.

Évaluation et amélioration de la fiabilité

Nous avons envisagé d'utiliser le « kappa » de Cohen, une mesure populaire de la fiabilité de l'intercodeur. Cependant, cette mesure suppose que les différents évaluateurs ont codé les mêmes éléments (par exemple, les mêmes paragraphes des mêmes textes). Dans notre cas, la difficulté provient du fait que les méthodes d'évaluation (nos unités d'enregistrement) ne sont pas clairement identifiées dans les documents de recherche en DS. Deux codeurs peuvent trouver différentes méthodes dans le même article, et la correspondance entre ces méthodes n'est pas connue à priori. Par conséquent, nous avons dû développer des métriques spécifiques.

Nous avons défini deux métriques de fiabilité de l'intercodeur, notées M_1 et M_2. Ces métriques sont calculées automatiquement pour chaque papier, pour chaque paire de codeurs (C_i, C_j). Ils sont basés sur un mappage un-à-un entre les méthodes d'évaluation identifiées par C_i et celles identifiées par C_j. Pour chaque article, l'algorithme considère le codeur qui a identifié le plus petit nombre de méthodes d'évaluation (entre C_i et C_j), et fait correspondre chaque méthode d'évaluation de ce codeur à une méthode d'évaluation de l'autre codeur. Parmi les mappages candidats, l'algorithme choisit la meilleure correspondance, c'est-à-dire celle qui minimise la distance moyenne entre les méthodes d'évaluation. La distance entre deux méthodes d'évaluation est la distance moyenne entre les caractéristiques des deux méthodes d'évaluation, pour chaque dimension de la taxonomie des méthodes d'évaluation. Les dimensions peuvent être plates ou hiérarchiques. Pour les dimensions hiérarchiques, nous appliquons la distance de généralisation [51]. Sur la base des mappages entre les méthodes d'évaluation, pour chaque papier et chaque paire de codeurs (C_i, C_j), la métrique M_1 est le pourcentage de méthodes d'évaluation cartographiées. Il mesure la mesure dans laquelle les codeurs Ci et Cj ont identifié le même nombre de méthodes. La métrique M_2 est la distance moyenne entre les méthodes d'évaluation mappées. Ces deux métriques sont de bons substituts à la mesure du kappa de Cohen, compte tenu des particularités de la situation de codage et de la nécessité d'établir à posteriori la correspondance entre les méthodes d'évaluation codées.

Après le codage du sous-échantillon, le pourcentage moyen calculé des méthodes d'évaluation cartographiées était de 59%, avec une distance moyenne de 0,31 entre les méthodes d'évaluation cartographiées. Ces scores assez faibles ont nécessité quelques changements dans le schéma de codage. Pour les réaliser, nous avons organisé un atelier auquel tous les codeurs ont participé. Sur la base de l'analyse des différences de codage, l'atelier a discuté des révisions de la taxonomie des méthodes d'évaluation, des ajouts aux heuristiques et aux exemples de codage, et des améliorations des heuristiques. En conséquence, la taille du document détaillant et illustrant les heuristiques a été doublée. Après un double codage de tous les articles avec le schéma de codage amélioré, le pourcentage moyen de méthodes d'évaluation cartographiées a atteint 71%, avec une distance moyenne de 0,21 entre les méthodes d'évaluation cartographiées. Considérant les progrès de la fiabilité de l'intercodeur, la complexité relative de la taxonomie utilisée pour le codage, et la spécificité du contexte de codage, nous considérons que ces valeurs sont raisonnables. Après le calcul de la fiabilité de l'intercodeur, les désaccords de codage ont été discutés et résolus entre chaque paire de codeurs, conduisant à une base de données de 402 méthodes d'évaluation.

Examen d'un sous-échantillon de documents et révision de la taxonomie

L'examen du sous-échantillon de dix articles (phase 4a sur la figure 1) a conduit à des modifications de la taxonomie (phase 4b). Ces changements découlent de la vérification des conditions d'arrêt, ou bien ont été suggérés par des recommandations de la littérature d'analyse de contenu. Actuellement, ces recommandations se chevauchent partiellement avec les conditions d'arrêt de la méthodologie de développement de la taxonomie. Par exemple, les auteurs de l'analyse de contenu suggèrent que les catégories devraient généralement s'exclure mutuellement [64] (les «catégories» dans le vocabulaire d'analyse de contenu correspondent aux «caractéristiques» dans le vocabulaire taxinomique présenté plus haut).

La première modification de la taxonomie a décomposé la caractéristique «test» en boîte noire et une autre blanche [26]. En raison de la rareté des tests en tant que méthode d'évaluation, nous n'avons pas fait cette distinction au départ. Cependant, l'analyse des désaccords entre les codeurs du sous-échantillon, ainsi que l'atelier dans lequel nous avons discuté de ces désaccords, a révélé l'ambiguïté du terme «test». En analyse de contenu, l'un des objectifs principaux est d'essayer le codage est la résolution des ambiguïtés. L'approche appliquée ici (découpage d'une catégorie en sous-catégories) est une technique de désambiguïsation classique.

La deuxième modification concerne la dimension «relativité de l'évaluation». Conceptuellement, «absolu», «relatif à l'absence d'artéfact» et «relatif à des artéfacts comparables» sont des notions qui apparaissaient comme s'excluant mutuellement, se conformant ainsi à la première condition d'arrêt dans la définition des taxonomies. Cependant, l'analyse du contenu du sous-échantillon a révélé des difficultés à distinguer entre les deux concepts : «absolu» et «relatif à l'absence d'artéfact». Nous avons donc décidé de ne conserver que les caractéristiques «absolues» et «relatives à des artéfacts comparables». L'artéfact n'est pas comparé aux autres, il englobe "par rapport à l'absence d'artéfact".

En définitive, à l'intérieur de la dimension «forme d'évaluation», le test du schéma de codage révèle l'ambiguïté du terme «analyse et raisonnement logique», ce qui semble suggérer que ces deux formes d'évaluation doivent toujours apparaître ensemble. Pour se conformer à la condition de fin que les caractéristiques dans une dimension doivent être collectivement exhaustives, nous avons renommé cette caractéristique comme «analyse ou raisonnement logique».

Résultats de l'analyse des documents de recherche en design

Cette section décrit la prochaine phase de la méthodologie de développement de la taxonomie (phase 5a). Dans cette phase, nous avons analysé le contenu de l'échantillon de 121 articles de recherche en conception. Les données résultant de l'analyse du contenu ont fourni la base pour l'analyse quantitative et à la compréhension de la pratique de l'évaluation des artéfacts SI. Nous présentons les résultats importants ci-dessous.

Lente maturation de la pratique dans la recherche en DS

Basé sur notre typologie d'artéfacts [46], nous avons calculé les fréquences par type et sous-type d'artéfact (la fréquence est le pourcentage d'articles contribuant à un artéfact de ce type ou sous-type). La conclusion la plus remarquable est la primauté des méthodes sur les modèles. Au sein des méthodes, les sous-types les plus courants sont l'algorithme (35%), la méthodologie (26%) et les lignes directrices (17%). Dans les modèles, toutes les fréquences sont inférieures à 15%. La primauté des méthodes sur les modèles est un signe que la pratique DSR mûrit.

Nous avons trouvé 402 méthodes d'évaluation des artéfacts dans l'échantillon d'articles (3,3 méthodes par article en moyenne). Le nombre moyen de méthodes d'évaluation par article sur une période a régulièrement augmenté depuis avril 2006 pour atteindre 3,9 au cours de la dernière période.

Pour évaluer la diversité des méthodes d'évaluation dans les différentes périodes, nous calculons le nombre de méthodes d'évaluation uniques pour chaque période de deux ans, soit toutes les combinaisons de (critère, technique d'évaluation, forme d'évaluation, participants secondaires, niveau d'évaluation, relativité de l'évaluation) dans les articles de cette période. "Unique" signifie que pour chaque période, nous comptons une seule fois les combinaisons apparaissant dans plusieurs articles de la période. La diversité des méthodes d'évaluation augmente constamment, passant de 39 dans la première période à 89 dans la dernière période. Si nous divisons le nombre de méthodes d'évaluation uniques par le nombre d'articles dans chaque période, ce rapport diminue entre la première et la deuxième période (2,44 à 1,85), mais il augmente ensuite régulièrement pour atteindre 2,70 au cours de la dernière période. La diversité accrue des méthodes d'évaluation est un autre signe que la pratique de l'évaluation des artéfacts SI mûrit. Nous complétons cette analyse en évaluant la diversité des méthodes d'évaluation pour les critères d'évaluation. À cette fin, nous calculons le nombre de méthodes uniques pour chaque critère, sur l'ensemble du périmètre des 121 articles. Le critère le plus diversifié dans les méthodes d'évaluation est l'*utilité* (trente-cinq méthodes d'évaluation uniques). Un examen plus approfondi des méthodes d'évaluation de ce critère révèle la variété des techniques d'évaluation (étude de cas, recherche-action, expérimentation en laboratoire ...) et des formes d'évaluation (mesurées, perçues à travers les items, qualitatives ...). Cette variété illustre l'influence de divers paradigmes (y compris la recherche qualitative et quantitative) sur l'évaluation des artéfacts dans les SI.

Le tableau 4 montre la fréquence de chaque forme d'évaluation (nombre et pourcentage d'articles utilisant chaque forme d'évaluation). L'analyse ou le raisonnement logique prédomine (par exemple, en démontrant l'efficacité des artéfacts avec des scénarios illustratifs, comme le confirmera l'analyse en grappes ci-dessous). Les métriques apparaissent dans 65 articles (54%). La recherche qualitative et quantitative enrichit les formes d'évaluation. Par exemple, l'utilité peut être évaluée sur la base de l'échelle de Davis [14]. La preuve formelle reste marginale dans l'évaluation des artéfacts.

Une analyse par approche d'évaluation révèle que les approches descriptives prédominent encore, malgré l'appel de Hevner et al. [26] pour les utiliser avec parcimonie : ces approches sont utilisées dans 52,9% des articles et sont suivies par des approches expérimentales (48,8%), des approches analytiques (19,8%) et des approches observationnelles ou participatives (18,2%). Ces résultats confirment également que dans le «débat réel versus laboratoire», «*la recherche en DS en laboratoire reste répandue*» [33, p. 174]: les approches empiriques dominent sur les approches observationnelles ou participatives. Une analyse détaillée par technique d'évaluation montre la prévalence de la simulation dans les approches expérimentales (37% des articles de notre échantillon) et des scénarios illustratifs dans les approches descriptives (48%). Les études de cas n'apparaissent que dans 13% des articles. Selon les définitions de notre système de codage, le terme «étude de cas» implique une situation réelle de résolution de problèmes. Ce terme est souvent abusé dans la littérature de recherche en DS, et un examen attentif des documents a révélé que de nombreuses «études de cas» étaient des scénarios illustratifs.

En résumé, nos données montrent que la recherche en DS, et plus précisément l'évaluation des artéfacts, mûrit lentement. Bien que la pratique de l'évaluation des artéfacts se diversifie, certaines tendances majeures subsistent (par exemple, la prévalence des approches descriptives et expérimentales).

Critères d'évaluation fréquemment évalués et inexplorés

Le tableau 5 montre la fréquence des critères d'évaluation (nombre d'articles évaluant chaque critère). Les critères les plus communs sont l'efficacité (N = 97), l'utilité (N = 42), faisabilité technique (N = 39), précision (N = 34), performance (N = 28), efficacité (N = 22), facilité d'utilisation, robustesse, évolutivité et faisabilité opérationnelle (N = 12). L'évaluation devrait établir que l'artéfact atteint son but et est utile. Ainsi, trouver l'efficacité, l'utilité et l'efficacité parmi les principaux critères n'est pas surprenant. Comme l'utilité, la facilité d'utilisation est un critère commun dans la littérature SI [14]. La faisabilité technique et opérationnelle est la catégorie la plus évaluée pour le critère « faisabilité ». Enfin, la précision, la performance, la robustesse et l'évolutivité sont souvent évalué avec des métriques. Des mesures facilement disponibles peuvent en partie expliquer fréquence d'évaluation de ces critères.

Parmi les trente-neuf critères, dix-sept (44%) ne sont jamais évalués dans notre échantillon. Plus précisément, une conclusion clé de cette étude est que l'éthique et les effets de bord ne sont jamais évalués. En ce qui concerne l'éthique, cela peut s'expliquer par l'absence des méthodes d'évaluation pour évaluer ce critère. En ce qui concerne les effets secondaires, ils peuvent seulement être pleinement évalués à long terme, ce qui n'est pas nécessairement compatible avec horizon temporel des publications de revues. Il est temps que la communauté de recherche en DS développe des approches pour évaluer pleinement l'impact organisationnel et sociétal des artéfacts. Ceci comprend des approches pour évaluer l'éthique [37] et les effets secondaires. De plus, l'alignement avec l'entreprise n'est jamais évalué dans l'échantillon, et la faisabilité économique est seulement évaluée une fois. Cela suggère une déconnexion partielle entre DSR et les concepts spécifiques à l'entreprise et les problèmes. Enfin, bon nombre des critères relatifs à la structure des artéfacts ne sont jamais évalués. La recherche en DS bénéficierait de l'adaptation des métriques de la littérature sur la qualité du logiciel et la qualité des données, par exemple, pour évaluer la simplicité structurelle des artéfacts.

Styles de composition dans l'évaluation des artefacts IS

Pour découvrir des motifs dans l'évaluation des artéfacts, nous avons regroupé les 402 méthodes d'évaluation de la base de données. En utilisant le langage et le système d'analyse statistique R [45], nous avons effectué un regroupement hiérarchique avec liens moyens, en ligne avec les suggestions de Hair et al. [24]. Nous avons préféré le regroupement hiérarchique - moyennes parce que nous ne connaissions pas le nombre de groupes à l'avance.

Nous avons cherché à découvrir les styles de composition les plus typiques, c'est-à-dire les motifs les plus typiques dans l'évaluation des artéfacts. Par conséquent, notre intérêt était sur les plus grands regroupements. Nous n'avions pas besoin de regroupements de tailles homogènes, d'où notre préférence pour lien moyen par rapport à la méthode de Ward.

Parce que les six dimensions des méthodes d'évaluation étaient symboliques, nous avions besoin d'une distance spécifique pour la classification hiérarchique. Nous avons réutilisé la distance définie pour la fiabilité évaluation (sous-section «Évaluation et amélioration de la fiabilité» ci-dessus). De le dendrogramme généré, nous avons choisi la solution de regroupement afin de minimiser l'augmentation de l'hétérogénéité tout en maximisant le pourcentage de méthodes d'évaluation classées dans des groupements de taille 20 ou plus (une taille de 20 correspond à environ 5% de la population totale des méthodes d'évaluation). Couper le dendrogramme avec un coefficient d'agglomération de 0,30 était le meilleur compromis. Cette solution comprend sept groupes de taille de 20 ou plus, représentant 78% de la population des méthodes d'évaluation. Pour les sept regroupements, la distance moyenne à l'intérieur des regroupements est de 0,2 ou moins. Tableau 6 montre les regroupements, avec leurs centroïdes.

Nous définissons un centroïde de regroupement comme une méthode d'évaluation (c.-à-d. une combinaison unique de caractéristiques pour les six dimensions de la taxonomie) de telle sorte que la distance moyenne avec les autres méthodes d'évaluation du regroupement est minime. Nous avons dérivé les noms des regroupements de leurs centroïdes et les méthodes d'évaluation les plus proches des centroïdes. Chaque cluster représente une composition style dans l'évaluation des artéfacts SI. Le style de composition le plus courant (le plus grand groupe de Tableau 6) est la démonstration. Ce style est typiquement utilisé pour démontrer, par analyse ou raisonnement logique, que l'artéfact fonctionne (efficacité, faisabilité technique), sur la base d'un scénario illustratif, en utilisant des exemples réels ou fictifs. Il ne nécessite pas de participants secondaires. Le deuxième style de composition le plus courant est la *simulation et la métrique d'étalonnage des artéfacts*: efficacité, précision, performance, robustesse ou évolutivité de l'artefact, est mesurée et comparée à celles résultant d'autres approches. De même que pour la démonstration, ce style ne nécessite pas de participants secondaires. L'évaluation de l'efficacité fondée sur la pratique établit généralement l'efficacité de l'artefact dans un cadre réel. Les techniques d'évaluation sont observationnelles ou participatives, et les praticiens participent à l'évaluation. Comme pour la méthode de démonstration, la forme typique d'évaluation est l'analyse ou le raisonnement logique. Le style de composition suivant est *simulation- et l'évaluation absolue métrique des artefacts*. Ce style est similaire à deuxième style et évalue les mêmes critères. La différence réside dans la relativité d'évaluation. Dans l'évaluation pratique de l'utilité ou de la facilité d'utilisation, le formulaire d'évaluation est généralement qualitatif. Laboratoire, évaluation par les étudiants de l'utilité mesure généralement la performance de la tâche des élèves en utilisant l'artéfact ou différent variations de l'artéfact. Enfin, l'analyse de complexité algorithmique étudie le temps ou la complexité de l'espace de l'artéfact (généralement, un algorithme), par une preuve formelle ou par analyse ou raisonnement logique (ce groupe comprend deux centroïdes).

Relations entre et dans le «quoi» et le «comment» de l'évaluation

Pour étudier les relations entre le «quoi» et le «comment» de l'évaluation, nous avons commencé par croiser les critères d'évaluation les plus fréquemment évalués (le «quoi») avec les cinq autres dimensions de l'évaluation (le «comment»). L'analyse du nombre de méthodes d'évaluation par critère et technique d'évaluation montre que pour l'efficacité, les scénarios illustratifs prédominent. Cela illustre encore le point en commun de la démonstration en tant que style de composition. Pour le critère utilité, les trois techniques les plus courantes sont (dans l'ordre): des scénarios illustratifs, des expériences de laboratoire et des études de cas. La précision, la robustesse et l'évolutivité sont généralement évaluées par simulation. La simulation est aussi la technique d'évaluation commune de la performance, à proportion égale des approches analytiques (analyse de la complexité). L'analyse par critère d'évaluation et forme d'évaluation montre que pour l'efficacité, le raisonnement analogique ou logique prédomine (il est utilisé dans 71 des 110 méthodes évaluant l'efficacité, et la combinaison de l'efficacité avec l'analyse ou le raisonnement logique est typique de la composition style). Les métriques sont utilisées dans 32 des 110 méthodes évaluant l'efficacité. Plus précisément, les articles contribuant à un modèle d'optimisation peuvent utiliser la valeur de la fonction objective comme mesure de ce critère. Les modèles mathématiques peuvent également être analysés formellement pour établir l'efficacité. En ce qui concerne l'utilité, les formes d'évaluation les plus courantes sont l'évaluation qualitative et les mesures. Les métriques d'utilité évaluent souvent les performances des tâches des utilisateurs. L'exactitude est généralement évaluée par la précision, le rappel ou la combinaison de ces mesures. Enfin, nos données suggèrent que le critère évalué influence la relativité de l'évaluation. Par exemple, l'évaluation de l'efficacité est généralement absolue. La forme et la relativité de l'évaluation sont liées (la plupart des 25 méthodes qui évaluent l'efficacité

relativement sont basées sur des mesures). Pour plus de précision, l'évaluation est principalement relative (comparaison de la précision avec des artéfacts comparables).

Pour explorer plus systématiquement les relations entre les différentes dimensions de l'évaluation, nous avons effectué des analyses statistiques approfondies en utilisant XLSTAT. Nous n'avons trouvé aucune corrélation significative entre les six dimensions de notre taxonomie des méthodes d'évaluation, confirmant ainsi que toutes les dimensions sont clairement distinctes. Comme mentionné ci-dessus, cela n'implique pas que ces derniers sont orthogonaux. Pour analyser plus finement les interdépendances entre les dimensions de l'évaluation, nous avons effectué une extraction de règles d'association avec R [23]. Les transactions étaient les 402 méthodes d'évaluation et les variables étaient les six dimensions de la taxonomie.

Le tableau 7 illustre quelques règles résultant de cette analyse (règles R1 à R4). Nous pouvons interpréter ces règles comme suit: *Généralement, l'évaluation du critère d'efficacité avec un scénario illustratif permet une évaluation absolue de ce critère* (règle R1). *Généralement, lorsque l'évaluation est effectuée avec des praticiens, en utilisant des exemples réels, cela aboutit à une évaluation absolue* (règle R2, suggérant que cette combinaison spécifique de participants secondaires et niveau d'évaluation peut ne pas être la plus appropriée pour une évaluation comparative). *Si aucun participant secondaire n'est disponible, la technique d'évaluation de la simulation peut être envisagée* (règle R3). *Si l'on souhaite évaluer les artéfacts de manière relativement précise, la simulation peut être une technique d'évaluation appropriée* (règle R4). En découvrant les règles, nous avons utilisé les valeurs de 0.8 et 0.1 pour la confiance minimum et le support minimum, respectivement. Ces seuils sont les valeurs par défaut dans la fonction R à-priori et sont couramment utilisés (par exemple, [13]). Avec ces contraintes de confiance et de soutien, nous avons trouvé plus de règles que celles montrées dans le tableau 7, qui illustre certaines des règles les plus pertinentes.

Pour compléter cette analyse, nous avons de nouveau effectué l'extraction des règles d'association, cette fois en incluant les types d'artéfacts apportés par chaque article. Parce que nous favorisons une vue systémique de l'évaluation des artéfacts SI et ne relions pas directement les types d'artéfacts aux méthodes d'évaluation, les transactions considérées pour cette analyse sont les articles (donc le nombre de transactions est passé de 402 à 121 résultats). Le but de l'analyse était de relier les types d'artéfacts avec les critères évalués et les techniques d'évaluation. Nous avons considéré chaque article comme un «panier» composé de tous les types d'artéfacts, de critères d'évaluation et de techniques d'évaluation de l'article. Le tableau 7 illustre deux des règles résultant de cette analyse: L'utilisation d'un scénario illustratif avec un exemple est un moyen possible d'évaluer l'efficacité d'une méthodologie (R5); Une simulation peut être recommandée pour évaluer la précision d'un algorithme (R6).

Évaluation

Dans cette section, nous évaluons la taxonomie des méthodes d'évaluation. Cette évaluation fait suite à l'examen de l'échantillon de 121 articles (phase 5a de la figure 1), détaillé dans la section précédente. Nous examinons les conditions d'arrêt et évaluons formellement la taxonomie, ce qui conduit à des révisions (phase 5b). Ces révisions concernent la hiérarchie des critères dans la taxonomie. Enfin, nous évaluons sommairement la taxonomie résultante.

Conditions d'arrêt et évaluation formative de la taxonomie

À ce stade, nous devons vérifier que tous les objets ou un échantillon représentatif d'objets ont été examinés. Dans notre cas, les objets sont les méthodes d'évaluation dans l'échantillon de 121

articles. Le processus de sélection de ces articles (expliqué dans la sous-section «Sélection des documents de recherche sur la conception» ci-dessus) assure la représentativité de l'échantillon. L'exigence selon laquelle au moins un objet doit apparaître sous chaque caractéristique de chaque dimension n'est pas satisfaite: comme mentionné précédemment, 17 des 39 critères de la hiérarchie des critères ne sont jamais évalués dans l'échantillon. Ces critères sont listés ci-dessous (nous les dénotons avec leurs catégories en cas d'ambiguïté): éthique, environnement / personnes / absence d'effets secondaires, alignement avec l'entreprise, environnement / organisation / absence d'effets secondaires, intégration dans l'architecture technique des SI, alignement avec Innovation informatique, environnement / technologie / absence d'effets secondaires, structure / simplicité, style, structure / cohérence, construction de surcharge, construction de redondance, construction d'excès, fonctionnalité, activité / cohérence, efficacité et modifiabilité. Nous considérerons cette liste de critères lors de la simplification de la hiérarchie.

Règle	Expression formelle de la règle	Confiance	Support
R1	{Évaluation Technique = Scénario descriptif / illustratif, Critère = objectif / atteinte des objectifs / efficacité} => {Relatif Ness_of_eval = Absolu}	1.00	0.13
R2	{Participant secondaire = Praticiens, Niv_de_evaluat= Instanciation / Exemple réel ou exemples} => {Relatif Ness_of_eval = Absolu}	0.94	0.20
R3	{Evalution_tech=Experimental/Simulation} => {Participant secondaire =aucun}	0.92	0.27
R4	{Form_of_eval=Quantitative/Measured, Relativeness_of_eval=Relative to comparable artifacts} => {Eval_technique=Experimental/Simulation}	0.87	0.15
R5	{Eval_technique=Descriptive/Illustrative scenario, Artifact_type=Instantiation/Example, Artifact_type=Method/Methodology} => {Criterion=Goal/ Goal attainment/Efficacy}	1.00	0.11
R6	{Criterion=Activity/Trustworthiness/Accuracy, Artifact_type=Method/Algorithm} => {Eval_technique=Experimental/Simulation}	0.94	0.14

Tableau 7. Relations entre et dans le "quoi" et le "comment" d'de l'évaluation: fouille de règle d'association

Dans la conceptualisation des dimensions et des caractéristiques de la taxonomie, nous avons considéré que la taxonomie était complète («complète», en termes de conditions d'arrêt), dans la mesure où elle reposait sur une revue de la littérature pertinente en science de conception. Nous devrions maintenant vérifier que les dimensions de la taxonomie incluent toutes les caractéristiques des objets, c.-à-d. des méthodes d'évaluation trouvées dans l'échantillon d'article. Plus précisément, nous devrions considérer la nécessité d'inclure des critères supplémentaires dans la hiérarchie des critères, dans le cas de critères évalués dans l'échantillon d'articles mais absents de la hiérarchie. Dans le système de codage (Annexe), les codeurs pourraient suggérer des critères supplémentaires. Lorsque les paires de codeurs se sont rencontrées pour discuter et résoudre les désaccords de codage, elles ont également confronté et discuté des critères additionnels suggérés par chaque codeur. Ceci a fourni la base pour décider des révisions de la hiérarchie des critères pour améliorer son exhaustivité, comme expliqué ci-dessous

Révision de la hiérarchie des critères

Nous avons utilisé les suggestions pour des critères supplémentaires et des discussions subséquentes pour améliorer l'exhaustivité de la hiérarchie. En ce qui concerne la simplicité, nous avons considéré les 17 critères ci-dessus comme des candidats potentiels à la suppression.

Amélioration de la complétude

Parmi les critères utilisés dans l'échantillon mais absents de notre hiérarchie, plusieurs critères (par exemple, la réutilisabilité) ne s'appliquaient qu'à des catégories spécifiques d'artéfacts. Nous n'avons pas ajouté ces critères à la hiérarchie, afin de préserver son applicabilité générale. Finalement, nous avons ajouté un critère et étendu la définition d'un autre critère. Nous avons ajouté la structure / la compréhensibilité. Ce critère apparaît dans la littérature scientifique de conception [34], mais nous ne l'avons pas inclus initialement, en le considérant comme un hyponyme de personnes / facilité d'utilisation. L'examen de l'échantillon a révélé que pour certains artéfacts, la compréhensibilité est clairement distincte de la facilité d'utilisation, d'où la nécessité de l'ajouter à la hiérarchie. En adaptant une définition du génie logiciel [29], nous définissons la compréhensibilité comme le degré de compréhension de l'artéfact, à la fois au niveau global et au niveau détaillé des éléments et des relations à l'intérieur de l'artéfact. En outre, nous avons étendu la définition de l'adaptabilité. Comme le révèle le codage des 121 articles, ce terme fait parfois référence à une réaction dynamique à des conditions environnementales changeantes. Puisque nous avions omis ce sens, nous avons complété la définition du critère comme suit: la facilité avec laquelle l'artéfact peut fonctionner dans des contextes autres que ceux pour lesquels il a été spécifiquement conçu, ou change selon les évolutions dans le contexte.

Amélioration de la simplicité

Nous avons simplifié la hiérarchie en supprimant certains des 17 critères non trouvés dans notre échantillon. Diverses raisons pourraient expliquer pourquoi un critère, suggéré par la littérature SI sur l'évaluation des artéfacts, n'a pas été trouvé dans l'échantillon de 121 articles. Cela nécessitait de prendre soin de décider quels critères devraient être conservés ou supprimés.

Comme mentionné ci-dessus, l'éthique et les effets secondaires sont cruciaux dans l'évaluation de l'impact des artéfacts. Nous devrions les garder dans la hiérarchie. La communauté de recherche en DS est invitée à définir des méthodes d'évaluation pour les évaluer. Les critères relatifs à la cohérence, l'alignement et l'ajustement sont liés. Ils apparaissent souvent dans la littérature de la science de conception et devraient être conservés dans la hiérarchie. Cet alignement avec les activités n'a pas été trouvé parmi les critères de notre échantillon est un signe de déconnexion partielle entre la recherche en DS et les préoccupations à l'échelle de l'organisation. Enfin, les critères de modification et de structure / simplicité, bien que non présents dans l'échantillon, s'appliquent à de nombreux types d'artéfacts, y compris les taxonomies (comme l'illustre cet article). Nous devrions les garder dans la hiérarchie.

Nous supprimons les autres critères de la hiérarchie. Malgré des suggestions précoces pour considérer le style (ou l'élégance) des artéfacts [26, 34], les chercheurs en sciences de conception ont montré peu d'intérêt pour ce critère. Aujourd'hui, l'évaluation de l'impact des artéfacts apparaît beaucoup plus critique que l'évaluation de leur style. En ce qui concerne la surcharge de construction, la redondance, l'excès et le déficit, ils devraient en fait être considérés comme des mesures de correspondance avec un autre modèle, au lieu de critères. Dans la catégorie «activité», le fait que la fonctionnalité n'ait jamais été évaluée suggère qu'il ne s'agit pas d'une catégorie spécifique. C'est un cas spécifique d'activité / complétude. Enfin, le codage des 121

articles a révélé la difficulté de distinguer l'efficacité de la performance. Nous supprimons l'efficacité de la hiérarchie des critères et redéfinissons les performances comme suit : le degré auquel l'artéfact accomplit sa fonction avec des contraintes de ressources données. Le temps et l'espace sont des cas spécifiques de ressources.

La figure 2 montre la hiérarchie révisée des critères. Pour conclure le processus de développement de la taxonomie, nous évaluons sommairement la taxonomie. Cela inclut l'évaluation de la simplicité pour évaluer l'impact des révisions à la hiérarchie des critères.

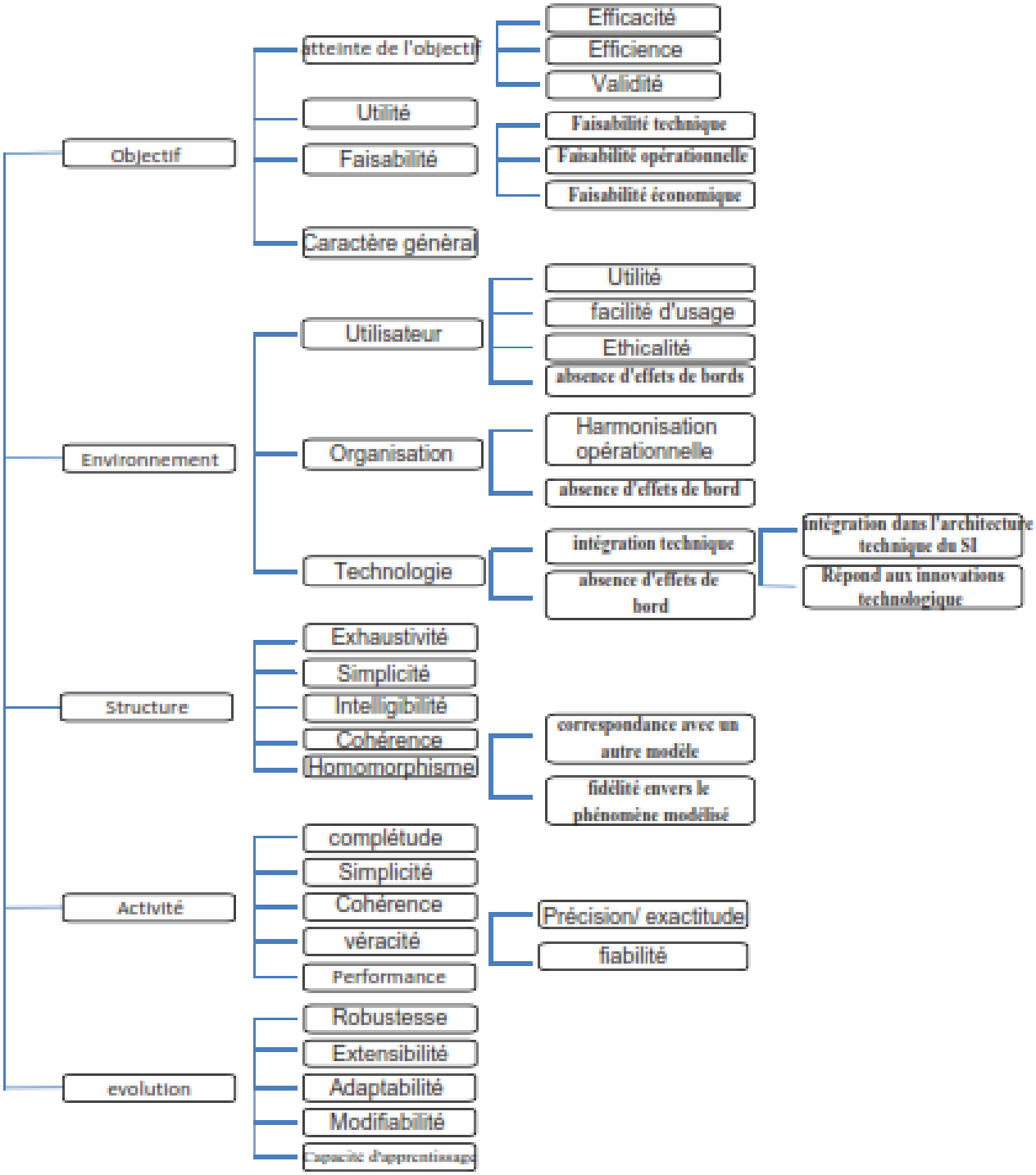

Figure 2 : hiérarchie finales des critères d'évaluation

Évaluation de manière cumulative de la taxonomie.

Tout d'abord il y a lieu de signaler que trois conditions d'arrêt subjectives sont à signaler au niveau de cet article : la taxonomie devrait être concise, robuste et explicative. «Concise» est l'équivalent du critère Structure/Simplicité dans notre hiérarchie de critères. L'application de la métrique du supplément B à cet article (en ligne) à la taxonomie finale donne une valeur de simplicité de 0,22 (contre 0,21 après la phase de conceptualisation). La simplicité de la hiérarchie des critères d'évaluation est de 0,27 (contre 0,26). Ainsi, les révisions de la taxonomie et plus particulièrement les révisions de la hiérarchie des critères n'ont au fond que légèrement amélioré la simplicité. La raison est que les ajouts de nouvelles caractéristiques ont partiellement compensé les simplifications. Ceci suggère également que notre métrique de la simplicité est conservatrice : même si l'effet net des révisions sur la hiérarchie était la suppression de cinq caractéristiques, l'impact était limité sur la simplicité calculée.

"Robuste" signifie que la taxonomie devrait fournir une différenciation entre les objets, c'est-à-dire l'expression de différences subtiles entre les objets dans la mesure où ces différences sont pertinentes. Ce critère est spécifique aux taxonomies et n'a pas d'équivalent dans notre hiérarchie de critères. Dans ce travail, nous devions représenter différents aspects de l'artificiel et du naturel dans les méthodes d'évaluation. La taxonomie permet l'expression de nuances dans la distinction entre évaluation artificielle et évaluation naturaliste. Par exemple, si l'on considère les dimensions «technique d'évaluation», «niveau d'évaluation» et «participants secondaires», les combinaisons de caractéristiques suivantes représentent une augmentation graduelle du caractère naturel (*scénario illustratif, exemple fictif ou exemples étudiants*), (*scénario illustratif, exemple réel ou exemples étudiants*), (*scénario illustratif, exemple réel ou exemples praticiens*) et (*étude de cas, exemple réel ou exemples praticiens*). La taxonomie permet également l'expression de nuances dans le «quoi» de l'évaluation en fonction de la hiérarchie des critères. Par exemple, la hiérarchie distingue cinq critères relatifs à l'évolution.

"Explicatif" signifie que la taxonomie devrait fournir des explications utiles sur la nature des objets classifiés. Pour évaluer l'utilité de la taxonomie, nous comparons cette recherche avec Venable et al. [58] (à notre connaissance, c'est le seul article du panier AIS dédié à l'évaluation en recherche en DS). Ces deux travaux semblent être complémentaires sur plusieurs aspects. Nous abordons les niveaux tactiques et opérationnels de l'évaluation des artéfacts SI, par rapport au niveau stratégique dans Venable et al. [58]. Nous nous concentrons sur le «quoi» et le «comment» de l'évaluation, par opposition à une vision de haut niveau du «pourquoi», du «quand», du «quoi» et du «comment». Nous définissons une hiérarchie de critères universels pour l'évaluation en recherche en DS, par opposition à la définition de critères universels qui nous semble plus problématique. Nous présentons des méthodes d'évaluation prototypiques (styles de composition) pour les artéfacts SI, versus des stratégies d'évaluation prototypiques.

Conclusion

Les contributions de cet article ont deux aspects. La première contribution est la taxonomie des méthodes d'évaluation des artéfacts SI. Cette taxonomie précise les dimensions relatives au «quoi» et au «comment» de l'évaluation, répondant ainsi aux questions QR1 et QR2. En ce qui concerne le «quoi» de l'évaluation, nous considérons le résultat d'un projet de recherche en DS comme un système d'artéfacts SI, et nous structurons la hiérarchie des critères d'évaluation en conséquence. Notre approche systémique fournit la vue holistique qui a manqué jusqu'à présent dans l'évaluation des artéfacts SI. Concernant le "comment", nous précisons les différentes dimensions des méthodes d'évaluation. Nous avons construit et évalué notre taxonomie selon le paradigme de la recherche en DS. Nous l'avons développé par une analyse conceptuelle de la littérature sur les sciences de conception, suivie d'une analyse empirique du contenu d'un échantillon d'articles de recherche sur la conception. Pour évaluer la taxonomie, nous avons appliqué certains critères issus de notre hiérarchie de critères.

Les résultats de l'analyse empirique de 121 articles de recherche en DS représentent notre deuxième contribution qui répond à la question QR3. Plus spécifiquement, cette analyse améliore la compréhension des relations entre les différentes dimensions de l'évaluation, identifie les motifs typiques (styles de composition) dans l'évaluation des artéfacts SI et met en évidence des critères inexplorés. Nous synthétisons les résultats de l'analyse empirique selon quatre lignes directrices d'évaluation pour les chercheurs en sciences de la conception.

Ligne directrice n°1: Envisager les styles de composition couramment utilisés dans l'évaluation des artéfacts SI. Les sept styles les plus courants sont (1) la démonstration, (2) la simulation et la métrique des artéfacts, (3) l'évaluation pratique de l'efficacité, (4) l'évaluation absolue basée sur la simulation et la métrique des artéfacts, (5) l'évaluation pratique de l'utilité ou de la facilité d'utilisation, (6) l'évaluation en laboratoire de l'utilité, et (7) l'analyse de complexité algorithmique. Ces styles peuvent être considérés comme des méthodes d'évaluation réutilisables. Le choix du style approprié dépend du critère à évaluer et des autres dimensions de l'évaluation.

Ligne directrice 2 : Générer de nouvelles méthodes d'évaluation de manière créative et pragmatique, en considérant les relations entre et dans le "quoi" et le "comment" de l'évaluation. Au-delà de la réutilisation des méthodes d'évaluation courantes, les chercheurs en sciences de conception sont également encouragés à générer de nouvelles méthodes. La variété des méthodes est déjà perceptible dans la pratique (par exemple, en ce qui concerne l'évaluation de l'utilité). Lors de la génération d'une méthode, les relations entre le «quoi» et le «comment» de l'évaluation, et entre les différentes dimensions du «comment», devraient être considérées. Le critère d'évaluation influence le choix de la méthode, plus précisément la technique d'évaluation et la forme et la relativité de l'évaluation. De plus, les règles d'association du tableau 7 illustrent certaines interdépendances entre ce qui est évalué (artéfacts SI composant le système évalué), critère d'évaluation, et les différentes dimensions dans le "comment" de l'évaluation (par exemple, la règle R4 concerne trois dimensions dans le "comment" de l'évaluation). Même si le "comment" de l'évaluation est généralement choisi en fonction du "quoi" (par exemple, les méthodes d'évaluation sont choisies en fonction des critères à évaluer), des considérations pragmatiques (par exemple, l'indisponibilité des participants secondaires) peuvent également influencer le choix des critères à évaluer. Ainsi, il y a plusieurs façons possibles d'interpréter les interdépendances entre le «quoi» et le «comment» de l'évaluation. En générant de nouvelles méthodes d'évaluation, la créativité est encouragée. Par exemple, pour évaluer les critères relatifs à la structure des artéfacts, les métriques d'ingénierie logicielle peuvent être adaptées.

Ligne directrice 3 : Étudier l'impact organisationnel des artéfacts SI. A en juger par notre échantillon d'articles, l'impact organisationnel, la mesure ultime du succès du SI [15], est négligé dans l'évaluation des artéfacts : dans la hiérarchie des critères, aucun critère relatif à l'environnement / organisation n'est évalué. L'accent est mis sur l'impact individuel (utilité). Le lien avec l'impact organisationnel est manquant. Comme Gill et Hevner [17] prétendent, et comme nos données le confirment, la recherche en DS a jusqu'ici surestimé l'utilité immédiate, au dépens d'un impact durable. L'évaluation de l'impact à long terme des artéfacts sur les organisations (y compris l'impact économique) devrait être encouragée.

Guideline 4 : Évaluer l'impact sociétal (y compris les considérations éthiques) des artéfacts IS, le cas échéant. L'impact sociétal devrait être une préoccupation majeure de l'évaluation des artéfacts [9,40]. Cependant, selon nos données, il n'est pas mesuré en pratique. Les chercheurs en SI peuvent innover en évaluant l'impact potentiel de leurs artéfacts sur la société. Cela nécessite le développement de nouvelles méthodes d'évaluation, par exemple, mesurer l'éthique et les effets secondaires.

Nos résultats, et les caractéristiques et dimensions de la taxonomie des méthodes d'évaluation, devrait être interprété à la lumière des limites de ce travail. Tout d'abord, cette recherche se concentre sur l'évaluation des résultats de recherche en DS. Elle ne considère pas l'évaluation du processus de recherche en DS lui-même. Deuxièmement, nous pourrions personnaliser la hiérarchie des critères pour refléter les spécificités de certains artéfacts, par exemple, les théories de conception. Troisièmement, au-delà du "comment" et du "quoi", notre taxonomie pourrait également considérer d'autres aspects de l'évaluation des artéfacts qui sont souvent implicites dans la recherche publiée, par exemple, le "pourquoi" [57]. Quatrièmement, dans la hiérarchie des critères, les trois catégories d'environnement [26] pourraient être enrichies pour inclure d'autres catégories d'acteurs, par exemple, la société en général. Cinquièmement, malgré la taille raisonnable de l'échantillon, cette étude devrait être reproduite dans des articles d'autres revues, par exemple, systèmes d'aide à la décision. Enfin, basé sur un plus grand échantillon, nous pourrions étudier les différences entre les principales revues en SI en ce qui concerne la pratique de l'évaluation des artéfacts. Ce sera l'objet de travaux futurs. Nous prévoyons également de développer des métriques, et plus généralement des méthodes d'évaluation, pour évaluer l'éthique et les effets secondaires des artéfacts.

Annexe : Système de codage

A. Codage global pour l'article
   1. L'artéfact de conception
      1.1 Type d'artéfact (choix multiples autorisés)
      1.2 Brève description des artéfacts
   2. Critères supplémentaires (le cas échéant) (évalués dans le document, mais absents de la hiérarchie des critères dans 6) (facultatif)

B. Codage pour chaque méthode d'évaluation
   3. Brève description de la méthode d'évaluation (pour une mesure, précisez le nom de la métrique)
   4. Technique d'évaluation. . .
   5. Forme d'évaluation. . .
   6. Critère évalué

Objectif

Atteinte de l'objectif

1. Efficacité : Le degré auquel l'artéfact atteint son but considéré de manière étroite, sans aborder les préoccupations situationnelles [8, 57].
2. Efficience : Le degré auquel l'artéfact atteint son but dans une situation réelle [8, 57].
3. Validité : La validité signifie que l'artéfact fonctionne correctement, c'est-à-dire qu'il atteint correctement son objectif [20].
4. Utilité : L'utilité mesure la valeur d'atteindre l'objectif de l'artéfact, c'est-à-dire la différence entre la valeur d'atteindre cet objectif et le prix payé pour l'atteindre [20, 69].

Faisabilité

5. Faisabilité technique : Évaluer, d'un point de vue technique, la facilité avec laquelle un artéfact proposé sera construit et exploité [5, 32].
6. Faisabilité opérationnelle : évalue dans quelle mesure la direction, les employés et les autres intervenants appuieront l'artéfact proposé, l'exploiteront et l'intégreront dans leur pratique quotidienne [5, 32].
7. Faisabilité économique : évalue si les avantages d'un artéfact proposé l'emporteraient sur les coûts de construction et d'exploitation de l'artéfact [5, 32].
8. Caractère générale : Fait référence à la portée de l'objectif de l'artéfact. Plus l'objectif est large, plus l'artéfact est général [3, 22].

Environnement

Utilisateur

9. Utilité : La mesure dans laquelle l'artéfact influence positivement la performance de la tâche des individus [14].
10. Facilité d'usage : La mesure dans laquelle l'utilisation de l'artéfact par les individus est sans effort [14].
11. Éthicalité : Le degré auquel l'artéfact est conforme aux principes éthiques.
12. Absence d'effets de bord : La mesure dans laquelle l'artéfact est exempt d'effets indésirables sur les individus à long terme [57]. Organisation

Organisation

13. Harmonisation opérationnelle : La congruence de l'artéfact avec l'organisation et sa stratégie [25].
14. Absence d'effets de bord : La mesure dans laquelle l'artéfact est exempt d'impacts indésirables sur l'organisation à long terme [57].

La technologie

Intégration technique

15. Intégration dans l'architecture technique des SI : degré auquel l'artéfact s'intègre dans l'architecture technique des SI de l'organisation.
16. Réponds aux innovations technologiques : la mesure dans laquelle l'artéfact utilise des TI novatrices [63].
17. Absence d'effets de bord : La mesure dans laquelle l'artéfact est exempt d'impacts indésirables sur l'architecture technique des SI de l'organisation à long terme [57]

Structure

18. Exhaustivité : degré auquel la structure de l'artéfact contient tous les éléments et relations nécessaires entre les éléments.
19. Simplicité : degré auquel la structure de l'artéfact contient le nombre minimal d'éléments et de relations entre les éléments [29].
20. Intelligibilité : L'élégance avec laquelle l'artéfact a été construit [26, 34].
21. Cohérence : Le degré d'uniformité, la standardisation et l'absence de contradiction entre les éléments de la structure de l'artéfact [29].

Homomorphisme

Correspondance avec un autre modèle : Le degré auquel la structure de l'artéfact correspond à un modèle de référence.

22. Surcharge de construit : La surcharge d'un construit se produit quand une construction dans la structure de l'artéfact est mappée à deux ou plusieurs constructions dans le modèle de référence [61].
23. Redondance de construits: la redondance d'un construit se produit lorsque deux ou plusieurs construits dans la structure de l'artéfact sont utilisées pour représenter une seule construction dans le modèle de référence [61].
24. Excès de construits : se produit quand un construit dans la structure de l'artéfact ne correspond à aucun construit dans le modèle de référence [61].
25. Déficit de construits : se produit quand un construit dans le modèle de référence ne mappe à aucun construit dans la structure de l'artéfact [61].
26. Fidélité aux phénomènes modélisés : degré auquel la structure de l'artéfact correspond à la réalité modélisée.

Activité

27. Complétude : Le degré auquel l'activité de l'artéfact contient tous éléments et relations nécessaires entre les éléments.
28. Fonctionnalité : La capacité de l'artéfact à fournir des fonctions qui répondent aux besoins exprimés et implicites lorsqu'il est utilisé dans des conditions spécifiées [29].
29. Simplicité : La mesure dans laquelle l'activité de l'artéfact contient le nombre minimal d'éléments et de relations entre les éléments [29].
30. Cohérence : Le degré d'uniformité, de standardisation et d'absence de contradiction entre les éléments de l'activité de l'artéfact [29].

Véracité

31. Précision : Le degré d'accord entre les sorties de l'artéfact et les résultats attendus [29].
32. Fiabilité : La capacité de l'artéfact à fonctionner correctement dans un environnement donné pendant une période de temps spécifiée [27].
33. Performance : Le degré auquel l'artéfact accomplit ses fonctions dans des contraintes de temps ou d'espace données. La vitesse et le débit (la quantité de sortie produite dans une période donnée) sont des exemples de contraintes de temps. L'utilisation de la mémoire est un exemple de contrainte d'espace [18, 29].
34. Efficacité : La maximisation du rapport entre les sorties et les entrées de l'artéfact.

Évolution

35. Robustesse : La capacité de l'artéfact à gérer les entrées invalides ou des conditions environnementales stressantes [29].
36. Extensibilité : La capacité de l'artéfact à manipuler des quantités croissantes de travail d'une manière gracieuse, ou à être facilement agrandie [6, 67].
37. Adaptabilité : La facilité avec laquelle l'artéfact peut fonctionner dans des contextes autres que ceux pour lesquels il a été spécifiquement conçu. Synonyme: flexibilité [29].
38. Modifiabilité : La facilité avec laquelle l'artéfact peut être changé sans introduire de défauts [29].
39. Capacité d'apprentissage : La capacité de l'artéfact à apprendre de l'expérience.

7. Participants secondaires. . .
8. Niveau d'évaluation. . .
9. Relativité de l'évaluation. . .