

# Understanding Benefits and Limitations of Unstructured Data Collection for Repurposing Organizational Data

Arturo Castellanos<sup>1</sup>(✉), Alfred Castillo<sup>2</sup>, Roman Lukyanenko<sup>3</sup>,  
and Monica Chiarini Tremblay<sup>4</sup>

<sup>1</sup> Baruch College (CUNY), New York City, NY, USA  
arturo.castellano@baruch.cuny.edu

<sup>2</sup> Cal Poly, San Luis Obispo, USA  
acast084@fiu.edu

<sup>3</sup> University of Saskatchewan, Saskatoon, Saskatchewan, Canada  
lukyanenko@edwards.usask.ca

<sup>4</sup> College of William and Mary, Williamsburg, VA, USA  
tremblay@fiu.edu

**Abstract.** With the growth of machine learning and other computationally intensive techniques for analyzing data, new opportunities emerge to repurpose organizational information sources. In this study, we explore the effectiveness of *unstructured data entry* formats in repurposing organizational data in solving new tasks and drawing novel business insights. Unstructured data accounts for more than 80% of the organizational data. Our research analyzes the implications of using unstructured data entry formats for propagation of organizational styles. We study this phenomenon in the context of case management in foster care. Using natural language processing and machine learning, we show that unstructured data formats foster entrenchment and propagation of individual organizational styles and deviations from the industry norms. Our findings have important implications both to theory and practice of business analytics, conceptual modeling, organizational theory and general data management.

**Keywords:** Systems analysis and design · Text mining · Stylometry · Unstructured data · Institutional theory · Case management

## 1 Introduction

Organizational data becomes a strategic resource for organizations. Effectively, these data can be aggregated to provide trends, plan, improve processes, support decision-making, or solve additional tasks by repurposing it. While some of these data are in structured and consistent form, organizational reports are often in unstructured format. IDC estimates that more than 80% of the enterprise data generated is unstructured [1].

Here, we define *unstructured data* as any document - clinical documentation, personal message, progress note, business report - that comprises primarily of unstructured text - with little or no predefined structure or meta data describing the content of the document. It is common to contrast unstructured data with *structured data* - such as information

stored in a spreadsheet or a database that follows a predefined structure or contains metadata describing the content of the stored information. Naturally, unstructured content does have internal structure, but its semantics needs to be discovered through additional processing (e.g., natural language processing) by a computer.

Despite the pervasiveness of unstructured data in organizations, traditional IS research offers limited guidance in understanding the implications of unstructured data-entry formats in decision-making – the alignment between the information needs of data consumers and that of data contributors. Data-entry refers to how these data is entered into a system (e.g., forms, templates or free-text fields).

One of the challenges of unstructured data formats is the inherent flexibility it gives to users when entering data into an information system—this may partially explain its popularity among data users. Users may deviate from the deep structure (“the meaning”) of the system by capturing different information in a field that was not originally intended for [2–5]. For example, in a study of an electronic patient record, physicians complained that the system was too “rigid” to capture the core reason of the patient’s visit. To overcome this perceived limitation physicians started to use a text field labeled as “conclusion” to enter such information and regarded it as a central field for subsequent patient’s visits [6, 7]. This is consistent with recent findings from the context of social media that suggest that imposing rigid structure when collecting information may result in users attempting to circumvent the structure by guessing or abandoning data entry entirely [8–11].

Despite the obvious benefits of unstructured information collection and its growing prevalence for organizational data capture and in social media environments, traditional research on conceptual modeling, systems analysis and information use mainly examined information collection in structured settings [12–14]. This creates a major gap in understanding of the limitations and benefits of the unstructured data collection, the gap where attempting to address in this and future work.

A better understanding of unstructured data collection is becoming increasingly important. Among other factors motivating our work is the on-going practice whereby organizations are repurposing data for business insight. This is possible due to increasing computational power and the availability of sophisticated analytical tools. For example, Tremblay, Berndt, Luther, Foulis and French [15] analyzed unstructured progress notes to predict falls in the elderly. Sørli, Perou, Tibshirani, Aas, Geisler, Johnsen, Hastie, Eisen, van de Rijn and Jeffrey [16] classified breast carcinomas based on variations in gene expression patterns and then correlate tumor characteristics to clinical outcome. Larsen and Bong [17] identified intellectual communities in the field of information systems and detected discordant naming practices of constructs (e.g., same term to refer to different phenomena or using different terms to refer to the same phenomena).

We focus on the effect of inferential utility in repurposing data. Our premise is that as people specialize they are more comfortable using domain-specific language. We demonstrate the relationship between inferential utility when repurposing unstructured electronic documentation and how institutionalization of practices need to be accounted for when designing more effective systems. Our goal is to demonstrate the implications of *unstructured data entry* on the ability of organizations to repurpose their existing tactical reports for strategic insight. We study this phenomenon in the context of case management in foster care (e.g., identifying cases of psychotropic drug use).

## 2 SafeKids

Our research is based on the triangulation of qualitative and quantitative evidence. Specifically, we draw on own experiences with the case of foster care in the United States. This setting allowed us to examine the issues related to unstructured data formats in a concrete and real scenario. This enabled us to produce qualitative insights into the nature of organizational reporting and the role of data formats. At the same time, we undertook systematic data collection from the foster care organizations we were observing to provide systematic analysis of the data and draw statistical inferences. We then returned to our qualitative understanding of the setting for additional corroboration and support of the quantitative evidence provided.

The organization that supplied the data for this paper is called SafeKids (name is anonymized) – an American non-profit corporation created by advocacy communities to oversee several Full-Case Management Agencies (FCMAs) that provide full case management services. Many of these cases include children at-risk of abuse and/or neglect. Failure to identify at-risk clients is highly problematic, because adverse outcomes can include serious adverse events—including death. Since data are often encoded in free-text form (e.g., reports, encounter notes, case notes, progress notes), we study the impact of different data-entry formats, in particular, when the goal is to repurpose these notes and use them for solving a different tactical need. We do so with a case study in which the tactical need is to identify children that are taking psychotropic medicines by analyzing the child’s case notes—as reported by caseworkers when visiting their homes.

In previous research Castillo et al. [18] hypothesized that by using these home-visit notes, which contained the child’s record and behavior (e.g. has signs of abuse and neglect, aggressive behavior), they could identify children taking psychotropic medication by training Statistical Text Mining (STM) classification models [19]. An interesting result was that models trained on individual FCMAs data had varying levels of classifications accuracy. This led us to ponder, if all agencies are not equal, did the writing style of each FCMA have an effect in improving the accuracy of our classification model? We turn to organizational theory and psychology theory to understand the underpinnings of flexibility in data-entry tasks.

## 3 Proposition Development

Institutions are organized and established by procedures that guide the actions of individuals [20]. Organizational activity (social and non-social) can become a pattern that is repeated by individuals in the organization. The rules, norms, and meanings arise in interaction and are preserved and modified by the behavior of individuals over time [21, 22]. In the absence of contextual change, actors are more likely to replicate scripted behavior, making these institutions persistent [23, 24]. Yet, behavior can evolve over time as a result of changing regulations and norms (e.g., solving an emergent tactical purpose or when solving wicked problems). The process of standardizing procedures among members of a population from these pillars is referred to as institutional isomorphism, which is triggered by coercive, normative, and mimetic

forces—constraining the ways in which individuals perform their activities [25]. This institutional isomorphism constrains the ways in which individuals perform their daily activities and cultivates expectations regarding the style of knowledge representations and production [25]. The concept of institutional isomorphism in organizational behavior theory leads to our first proposition:

**Proposition 1 (homogeneity):** *Data collected using unstructured-data-entry formats become homogenous within organizational units. This homogeneity is more prominent within the same organizational unit.*

The effectiveness on their decision-making is tied to the information at hand to solve such tactical purpose. This *data* homogeneity would suggest the potential for organizations to adopt standard practices in how they collect and use the information to solve a tactical need. Institutional features of organizational environments, however, can shape the actions actors take (e.g., the level of detail—specificity or focus—at which they input the information into the IS). Moreover, because of institutionalization of practice, notes from one organizational unit are similar to one another and less similar than notes from different organizational units. More importantly for the organization, is to find a way to assess the effectiveness of these unstructured notes in solving a task.

Free-text data collection’s flexibility implies that the level of detail of case notes can vary across individuals across organizational units. We turn to theories from psychology to discuss the tradeoff between generalization and specification in data collection. According to psychology, categories support vital functions of an organism via *cognitive economy* and *inductive inference* [26–30]. Cognitive economy is achieved by maximally abstracting from individual differences among objects and then grouping objects in categories of larger scope [28, 31, 32]. These categories improve the ability of a person to accurately predict features of instances of a category. The trade-off between these competing functions is considered one of the defining mechanisms of human cognition and behavior [27, 33]. According to cognitive theories and theories of classification, categories (which can be represented as a class in the IS) provide cognitive economy and inferential utility, enabling humans to efficiently store and retrieve information about phenomena of interest [27, 30]. In a free-text interface these categories are not fixed and are chosen by the individual entering the data into the system.

Lukyanenko, Parsons and Wiersma [9] suggests that in a free-form data entry task, non-experts will classify more accurately at a general level than at a more specific level. When we collect structured data the level of specificity is fixed at the time of system design. Users entering unstructured data, on the other hand, can adjust to their level of specificity—by being more or less detailed [34]. Since specificity results from expertise, unstructured data collection can capture expertise better, which in turn may lead to better performance by having relevant information to support decision-making (e.g., repurposing existing data). We suggest that organizations can foster effective unstructured-data-entry practices that could result in richer data collection. We do so through the following propositions:

**Proposition 2 (Inferential utility and repurposing):** *Unstructured-data-entry formats can help shape effective data-entry practices in solving well-defined needs.*

**Proposition 2a:** *Higher levels of specificity in the unstructured data collected leads to increased inferential utility.*

**Proposition 2b:** *Higher levels of specificity in the unstructured data collected facilitate repurposing data for other tasks.*

The goal of the proposed design propositions is to understand the subtleties of unstructured-data-entry electronic documentation to design more effective information systems [35, 36]. The propositions enable designers to reflect on the effect of institutional practices in user generated electronic documentation. In the following sections we evaluate the propositions and provide a discussion, conclusions, and areas for future research.

## 4 Evaluation of Propositions

To evaluate Proposition 1 we use Stylometry, a particular application of text mining. To evaluate Proposition 2, we use text-mining techniques to analyze differences in the text authored by different case workers.

### Proposition 1: Homogeneity of Data

Some researchers have argued that an author's style is comprised of a limited number of distinctive features inherent to the author, neglecting the content/context-dependency of the writing [37]. Stylometric analysis is an application of text mining that uncovers metadata from the documents and allows for statistical comparisons of these metadata as a proxy for "style". We use SAS Text Miner 9.4 to predict, based on the text in the case note, to which FCMA a particular case note belongs.

Our training set consists of all the case notes from 795 children from three agencies assigned to a mutually exclusive train and test set. We train a classification model that has the case note text and our target variable—the FCMA from which that note is coming from (e.g., FCMA A – 336 children in total, FCMA B – 213 children in total, and FCMA C – 246 children in total).

We create individual models for each FCMA and we evaluate the performance of the predictive models using commonly accepted metrics: recall, precision, and F-measure (see Table 1). Our results show that despite organizations having established guidelines of reporting, employees adopt new guidelines that become norms over time. This is reflected in how different organizational units are consistent in the way they encode home-visit notes. We also introduce the idea of organizational stylometry. To our knowledge, the use of stylometry at the population level (where many contributors to a body of text) has yet to be explored.

The results of the analysis show that by analyzing a particular case note we can predict, with a high degree of certainty, the authoring FCMA of that case note (see Table 2). These results show that each organization has its unique style, which is consistently used by its caseworkers. Based on these results we can confirm Proposition 1 that *institutional factors establish data entry practices that result in data that is similar within organizational units.*

**Table 1.** Evaluation metrics

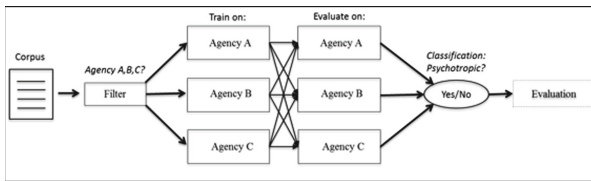
Precision (P)	Recall (R)	F-measure
$P = \frac{TP}{TP + FP}$	$R = \frac{TP}{TP + FN}$	$F = \frac{2(P * R)}{P + R}$

**Table 2.** Evaluation metrics across agencies

Agency	Precision (%)	Recall (%)	F-measure (%)
Agency A	76.99	75.65	76.31
Agency B	79.81	76.15	77.94
Agency C	76.74	81.15	78.88

### Proposition 2: Inferential Utility and Repurposing

To evaluate Proposition 2, we use an inductive (classification) text mining technique. First, an expert case manager provides a gold standard with labeled instances. Case notes are labeled “Yes” (uses psychotropic medication) or “No” (no use of psychotropic medication), depending on whether the child is taking psychotropic medication or not. We create individual models for each FCMA (A, B, and C) and we evaluate each within its own organizational unit (intra) and across organizational units (inter) (see Fig. 1). For each organizational unit, we assign a random sample into a training set containing 70% of the cases and a test set containing the remaining 30% of the data [15, 18]. We use SAS Text Miner 9.4 to evaluate the performance of each of the models and all the permutation comparisons across organizational units.



**Fig. 1.** Intra and Inter-agency data mining process

There is no standard definition of what a substantial difference in F-measure improvement should be. In the field of information retrieval a 5% performance improvement is considered a substantial improvement [38, 39]. The z-test for proportions evaluates the statistical difference between two population proportions  $p_1$  and  $p_2$  [40, 41]. To test the difference between proportions we compute the following:

$$z_{proportions} = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (1)$$

We evaluated each FCMA by comparing the performance when tested with data from the same organizational unit (intra-FCMA) and compared to models that use data from other organizational units (inter-FCMA). We highlight in bold any statistically significant differences for precision and recall using a z-test for proportions (two-tailed test at the 95% confidence level). We consider the difference in F-measure as substantial if the difference between F-measures is more than 0.05 and the difference in precision or recall is statistically significant (determined using the z-test for proportions and highlighted in bold and with a \* symbol) [38].

Table 3 shows that the differences in F-measure are substantial in five out of the six pairs. The results show that two of the agencies (FCMA A and FCMA C) consistently perform better in classifying cases of psychotropic drug use. This shows that unstructured data entry formats may result in differences in how information is collected across different organizational units in the organization. Institutional theory helps explain how institutional factors shape practices by individuals in different organizational units, and how these practices can become stable over time and adopted by other individuals, making practices persistent. This validates our first proposition that *data collected using unstructured-data-entry formats become homogeneous within organizational units*.

Common NLP tools include document tokenizing, stemming, parts-of-speech tagging, noun group extraction, applying stop lists, entity identification, and multiword terms handling [42]. The document is parsed and tagged based on the syntactical relationship between terms –based on the position in a sentence and rules of grammar [43].

**Table 3.** Difference between proportions for precision (P) and Recall (R)

Train	Evaluation	Precision	Recall	F-Measure
FCMA A	FCMA A	78.57	70.97	74.58
	FCMA B	65	52.7	<b>58.21</b>
	FCMA C	<b>31.94</b>	<b>30.67</b>	<b>31.29*</b>
	Z-Value (FCMA A-FCMA B)	1.2858	1.7303	
	Z-Value (FCMA A-FCMA C)	<b>4.2082</b> (p<0.01)	<b>3.9911</b> (p<0.01)	
FCMA B	FCMA B	46.15	54.54	50
	FCMA A	45.59	30.69	<b>36.69*</b>
	FCMA C	32	21.33	<b>25.6*</b>
	Z-Value (FCMA B-FCMA A)	0.1377	<b>2.1261</b> (p<0.05)	
	Z-Value (FCMA B-FCMA C)	1.1864	<b>3.0229</b> (p<0.01)	
FCMA C	FCMA C	64.71	50	56.41
	FCMA A	<b>33.33</b>	<b>16.83</b>	<b>22.37*</b>
	FCMA B	59.26	<b>21.62</b>	<b>31.68*</b>
	Z-Value (FCMA C-FCMA A)	<b>2.2762</b> (p<0.05)	<b>3.3621</b> (p<0.01)	
	Z-Value (FCMA C-FCMA B)	0.3613	<b>2.5992</b> (p<0.01)	

The aim is to convert human language into formal representations computers can manipulate, including part-of-speech tagging (POS), POS sequences, or n-gram models [42, 44, 45]. Because authors do not always follow grammatical rules, the complexity of multiple meanings for words, and the domain specific use of vocabulary may require some additional considerations.

Results show that there is no statistical significant difference between the full model and the model that has no Part-of-Speech and Noun Group features but does have a term weighting scheme (Term Frequency, Term Weight). Consistent with previous research, the terms used are a more salient factor of prediction compared to the language structure of a case note. Human language is subtle, with many unquantifiable yet salient qualities. Users with different levels of expertise tend to produce information that differs in quality and level of abstraction. For example, within the category “taking medication”, a conceptual hierarchy can be the following: (a) medication (b) psychotropic medication (c) Lisdexamfetamine (d) Vyvanse, which goes from the most general (a) to the most specific (d). Knowing a child is taking Vyvanse (d) gives more information than just knowing a child is taking medication (a).

We assess language use (in terms of structure and meaning of the case notes) by including/excluding natural language processing (NLP) features.

A case note authored by Agency A “Takes mg of vyvanse by mouth once a day [...] she has to call the doctor to schedule for a refill” has a confidence of 0.953 of being psychotropic drug use case. Whereas the following case note authored by Agency C “child has an immune system medical condition that requires many medications to keep her healthy” has a confidence of 0.594 of being a case of psychotropic drug use –as measured by the singular vector decomposition scores. The results of the analysis show that higher levels of specificity in the data collected leads to increased inferential utility, which can ultimately help the organization solve unanticipated tasks using these data. This validates Proposition 2 that *higher levels of specificity in the unstructured data collected leads to increased inferential utility, which can in turn be leveraged to repurpose data for a different task it was originally designed for*.

In the final section we discuss the implications of our findings for theory and practice.

## 5 Implications for Research and Practice

Our findings have important implications both to theory and practice of conceptual modeling, unstructured information collection, text mining and business analytics, and general organizational data management.

We believe our work is timely. Traditional information systems research has been concerned with finding similar elements in highly structured data sets [46] and the study of unstructured data sources is a relatively recent active stream of work. At the same time, unstructured data sources continue to grow in prominence fueled by the explosive growth in social media and online content production which tends to be text-based. Our work aims to provide both theoretical and practical insight into the nature of unstructured information. The arguments and findings of our work are thus applicable to user generated content settings and as we as our context of corporate



unstructured data [13]. Indeed, researchers continue to call for novel approaches to structure user generated content to make it more consistent and usable in organizational analysis [11]. Our work has strong potential to contribute to the efforts to make user generated content more usable by increasing its potential for reuse. In the future, we hope to extend our work to the area of user generated content (specifically, crowd-sourcing) to address the issue of repurposing it for unanticipated insights.

In our paper, we show that we can reliably detect organizational styles. This insight can be used to improve organizational processes and foster more effective data reuse. First, our research suggests that the data-entry formats of the information system can highlight the existence of different organizational styles across organizational units. Second, our research suggests that the flexibility of free-form data entry motivates individuals to stay truthful to their organizational unit's reporting expectations. This highlights the trade-off between different data-entry formats and the data collected by the organization.

Our results demonstrate the role of the level of specificity in enabling unanticipated insights. The results of this study can be generalized to other domains and can provide insight to effective system design—the effect of particular designs (that are more/less flexible). In a fully structured scenario, the user is guided by the interface on what needs to be reported. In a semi-structured scenario, pre-established templates guide data entry but allows for some deviation by the user to input something not related to a particular template. In an unstructured scenario (e.g., free-form), the individual has the liberty to enter data, which is typically defined by the organization (e.g., business processes, training).

Our research encourages experts to be as specific as they can while allowing non-experts to input information at a more general level. Higher specificity, however, requires higher expertise. Thus, it may hinder collaboration from non-experts. Future work should focus on how these different data-entry formats may preclude the collection of valuable information (leading to information loss) when both novices and experts contribute to the system. Previous research have shown that limiting data-entry to experts can preclude the input of valuable information from non-experts and can lead to data accuracy problems [9].

Our results are consistent with psychology research in that the level of specificity of the information limits the applications for which that data can be used. In an environment where individuals have a similar level of expertise based on their background and training, it is preferable for them to be more specific when they enter the data into the system (e.g., the child is taking 5 mg of Adderall provides more information than just saying the child is taking medication). A practical implication to this is that depending on whether the individuals looking at the text are a non-experts vs. experts, the individual writing the text can choose to contribute beyond what he believes is the information required for the reader. This allows for increased inferential utility that can prove beneficial when dealing with unanticipated uses of the data.

Our study also provides guidance of the implications of choosing how data entry formats of a system are designed—and what is it that they would like to capture from their users. To the best of our knowledge, Authorship Analysis had only been done at the individual level. We extend this analysis for authorship identification at the group level (e.g., identifying the authoring organizational unit of a body of text). This can be

used by organizations to assess the consistency of data-entry practices in an organization and can be extended by using analytical techniques to create dimensions of categories these documents fall into or metrics that relate to reliability of the data.

The tension between data collection at different levels of granularity further suggests exciting new opportunities at the intersection of conceptual modeling and data analytics. Conceptual modeling research has long studied the nature of content aggregation, part – whole relationships and the general ontological assumptions behind data collection [47–49]. These can become valuable sources of guidance for innovative analytics approaches aiming at drawing inferences from data collected at different levels of analysis. We hope to pursue this work in the future.

## 6 Limitations

This study is not without limitations. There is a threshold for the classification models accuracy that is directly related to the quality of the data in the gold standard. For instance, psychotropic medication was attributed to the foster home and not the child. If a foster home has multiple children and one was taking psychotropic medication, all of these children would appear as taking psychotropic medication and vice-versa. This is a limitation that introduces biases in the classification models. Moreover, we did not take into account time windows (e.g., a kid that was prescribed psychotropic medication is no longer taking that medication). However, this does not undermine the goal of our work, which is to understand the relationship of data-entry practices in repurposing data. Future work should focus in providing a method to evaluate when using data in the aggregate is justified as opposed to highlighting meaningful segments for separate analysis.

## References

1. Gantz, J., Reinsel, D.: Extracting value from chaos. IDC Iview **1142**, 1–12 (2011)
2. Boudreau, M.-C., Robey, D.: Enacting integrated information technology: a human agency perspective. *Organ. Sci.* **16**, 3–18 (2005)
3. Wand, Y., Weber, R.: On the deep structure of information systems. *Inf. Syst. J.* **5**, 203–223 (1995)
4. DeSanctis, G., Poole, M.S.: Capturing the complexity in advanced technology use: adaptive structuration theory. *Organ. Sci.* **5**, 121–147 (1994)
5. Burton-Jones, A., Grange, C.: From use to effective use: a representation theory perspective. *Inf. Syst. Res.* **24**, 632–658 (2012)
6. Berg, M., Goorman, E.: The contextual nature of medical information. *Int. J. Med. Inform.* **56**, 51–60 (1999)
7. Berg, M.: Implementing information systems in health care organizations: myths and challenges. *Int. J. Med. Inform.* **64**, 143–156 (2001)
8. Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., Cox, A.L.: Designing for dabblers and deterring drop-outs in citizen science. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2985–2994. ACM (2014)

9. Lukyanenko, R., Parsons, J., Wiersma, Y.F.: The IQ of the crowd: understanding and improving information quality in structured user-generated content. *Inf. Syst. Res.* **25**, 669–689 (2014)
10. Van Kleek, M.G., Styke, W., Karger, D.: Finders/keepers: a longitudinal study of people managing information scraps in a micro-note tool. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2907–2916. ACM (2011)
11. Lukyanenko, R., Parsons, J., Wiersma, Y., Wachinger, G., Huber, B., Meldt, R.: Representing crowd knowledge: guidelines for conceptual modeling of user-generated content. *J. Assoc. Inf. Syst.* **18**, 2 (2017)
12. Jabbari Sabegh, M.A., Lukyanenko, R., Recker, J.C., Samuel, B., Castellanos, A.: Conceptual modeling research in information systems: what we now know and what we still do not know (2017)
13. Burton-Jones, A., Volkoff, O.: How can we develop contextualized theories of effective use? A demonstration in the context of community-care electronic health records. *Inf. Syst. Res.* (2017)
14. Lukyanenko, R., Parsons, J.: Information quality research challenge: adapting information quality principles to user-generated content. *J. Data Inf. Qual. (JDIQ)* **6**, 3 (2015)
15. Tremblay, M.C., Berndt, D.J., Luther, S.L., Foulis, P.R., French, D.D.: Identifying fall-related injuries: text mining the electronic medical record. *Inf. Technol. Manage.* **10**, 253–265 (2009)
16. Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S.S.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98**, 10869–10874 (2001)
17. Larsen, K., Bong, C.H.: A tool for addressing construct identity in literature reviews and metaanalyses. *MIS Q.* **40**, 529–551 (2016)
18. Castillo, A., Castellanos, A., Tremblay, M.C.: Improving case management via statistical text mining in a foster care organization. In: Tremblay, M.C., VanderMeer, D., Rothenberger, M., Gupta, A., Yoon, V. (eds.) *DESRIST 2014*. LNCS, vol. 8463, pp. 312–320. Springer, Cham (2014). doi:[10.1007/978-3-319-06701-8\\_21](https://doi.org/10.1007/978-3-319-06701-8_21)
19. Luther, S., Berndt, D., Finch, D., Richardson, M., Hickling, E., Hickam, D.: Using statistical text mining to supplement the development of an ontology. *J. Biomed. Inform.* **44**, S86–S93 (2011)
20. Jepperson, R.L.: Institutions, institutional effects, and institutionalism. *New Institutionalism Organ. Anal.* **6**, 143–163 (1991)
21. Giddens, A.: *Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis*. University of California Press, Berkeley (1979)
22. Sewell Jr., W.H.: A theory of structure: Duality, agency, and transformation. *Am. J. Soc.* **98**, 1–29 (1992)
23. Hughes, E.C.: The ecological aspect of institutions. *Am. Sociol. Rev.* **1**, 180–189 (1936)
24. Barley, S.R., Tolbert, P.S.: Institutionalization and structuration: Studying the links between action and institution. *Organ. Stud.* **18**, 93–117 (1997)
25. DiMaggio, P.J., Powell, W.W.: The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *Am. Soc. Rev.* **48**(2), 147–160 (1983)
26. Lakoff, G.: *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago (1987)
27. Roach, E., Lloyd, B.B., Wiles, J., Rosch, E.: *Principles of categorization* (1978)
28. Smith, E.E., Medin, D.L.: *Categories and Concepts*. Harvard University Press, Cambridge (1981)
29. Smith, E.E.: Concepts and thought. In: *The Psychology of Human Thought*, p. 19 (1988)

30. Parsons, J.: An information model based on classification theory. *Manage. Sci.* **42**, 1437–1453 (1996)
31. Fodor, J.A.: *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press, Oxford (1998)
32. Murphy, G.L.: *The Big Book of Concepts*. MIT Press, Cambridge (2004)
33. Corter, J., Gluck, M.: Explaining basic categories: feature predictability and information. *Psychol. Bull.* **111**, 291–303 (1992)
34. Lukyanenko, R., Castellanos, A.: Introducing information gradient theory. In: *Breakthroughs and Emerging Insights from Ongoing Design Science Projects: Research-in-progress papers and poster presentations from the 11th International Conference on Design Science Research in Information Systems and Technology (DESIST 2016)* 2016, St. John, Canada, 23–25 May (2016)
35. Walls, J.G., Widmeyer, G.R., El Sawy, O.A.: Building an information system design theory for vigilant EIS. *Inf. Syst. Res.* **3**, 36–59 (1992)
36. Eisenhardt, K.M.: Building theories from case study research. *Acad. Manag. Rev.* **14**, 532–550 (1989)
37. De Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. *ACM Sigmod Rec.* **30**, 55–64 (2001)
38. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst. (TOIS)* **23**, 103–145 (2005)
39. Sparck Jones, K.: Automatic indexing. *J. Doc.* **30**, 393–432 (1974)
40. Kachigan, S.K.: *Statistical Analysis: An Interdisciplinary Introduction to Univariate & Multivariate Methods*. Radius Press, New York (1986)
41. Fleiss, J.L., Levin, B., Paik, M.C.: *Statistical Methods for Rates and Proportions*. Wiley, New York (2013)
42. Manning, C.D., Schütze, H.: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (1999)
43. Berry, M.W., Castellanos, M.: Survey of text mining. *Comput. Rev.* **45**, 548 (2004)
44. Abbasi, A., Chen, H.: CyberGate: a design framework and system for text analysis of computer-mediated communication. *Mis Q.* **32**(4), 811–837 (2008)
45. Holmes, D.I.: The evolution of stylometry in humanities scholarship. *Literary Linguist. Comput.* **13**, 111–117 (1998)
46. Batini, C., Lenzerini, M., Navathe, S.B.: A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv. (CSUR)* **18**, 323–364 (1986)
47. Shanks, G., Tansley, E., Nuredini, J., Tobin, D., Weber, R.: Representing part-whole relationships in conceptual modeling: an empirical evaluation (2002)
48. Evermann, J., Wand, Y.: Towards ontologically based semantics for UML constructs. In: Kunii, H.S., Jajodia, S., Sølvberg, A. (eds.) *ER 2001. LNCS*, vol. 2224, pp. 354–367. Springer, Heidelberg (2001). doi:[10.1007/3-540-45581-7\\_27](https://doi.org/10.1007/3-540-45581-7_27)
49. Wand, Y., Storey, V.C., Weber, R.: An ontological analysis of the relationship construct in conceptual modeling. *ACM Trans. Database Syst. (TODS)* **24**, 494–528 (1999)

Information Systems: Research, Development,  
Applications, Education

10th SIGSAND/PLAIS EuroSymposium 2017, Gdansk,  
Poland, September 22, 2017, Proceedings

Wrycza, S.; Maślankowski, J. (Eds.)

2017, XII, 153 p. 34 illus., Softcover

ISBN: 978-3-319-66995-3