What Makes Data Possible? A Sociotechnical View on Structured Data Innovations

Aleksi Aaltonen
Temple University
aleksi@temple.edu

Esko Penttinen
Aalto University
esko.penttinen@aalto.fi

Abstract

Drawing from the theory of digital objects, this paper examines the distinction between structured and unstructured data as carriers of facts. We argue that data do not 'have' a structure but are made by a structure that confers data their capacity to represent contextual facts. We employ a case vignette involving XBRL (eXtensible Business Reporting Language) and its use in statutory financial reporting to illustrate and explore the sociotechnical nature of data and to describe what we call data innovations: new valuable ways to render phenomena as data. We find that data structure is best viewed as a matter that is relative to a purpose in a context. Theorizing data from a sociotechnical perspective could evolve to provide, in effect, the material science of digital economy.

1. Introduction

The quantities of data generated in the digital economy are growing at a prodigious rate [19], and many academics and practitioners increasingly view data as the new 'oil' for the post-industrial society [36]. In consequence, firms and entire industries have awakened to the fact that they must be able to harness this new digital resource to remain competitive. However, most of the data are generated in so-called 'unstructured' form that limits their applicability for various purposes. A deluge of human-generated messages and documents, photos, video and audio recordings, and social media contents sweeps into information systems every day, and even machine-generated data can be often poorly structured beyond its immediate usage.

The difference between structured and unstructured data is seemingly easy to grasp. In general, structured data are recorded as well-defined fields that correspond to distinct variables, whereas unstructured data, such as natural language writings, consist of a mishmash of semantic entities that can differ from an observation to another and it may not even be clear what constitutes a separate observation in unstructured data. Analytics, which is the primary means by which value is extracted from data, usually assumes the availability of sufficiently structured data. If structured data are not available, data mining and machine learning techniques can sometimes be used to reconstruct a latent structure hidden in seemingly unstructured data. For example, one might employ a topic model to represent the text of product reviews as feature vectors and then classify the reviews on the basis of the vectors, thus rendering the review content amenable to analytical operations.

However, the clarity of the distinction between structured and unstructured data starts to break down upon closer inspection. Those which are considered unstructured data in one setting can function as structured data in another context. For instance, a bitmap image of a company's annual results is unstructured data in the sense that the revenue, profit, and other financial information in the image are computationally inaccessible to further financial analysis. At the same time, the data in the bitmap can be processed by an image compression algorithm that identifies visual structures in the data and reduces the file size without degrading its image quality. In fact, we will show that there can be no completely unstructured data from a computational perspective; all digital data are ultimately structured as binary distinctions [40], which must be accompanied by some rudimentary knowledge on how to combine the distinctions into higher-level entities such as characters by using character encodings, or pixels of a bitmap image, etc.

In this paper, we problematize data structure as an essentially relational matter. Data can be variously structured with respect to different purposes but, to be perceived as data, digital inscriptions must be

embedded in a structure that allows contextualizing and making sense of their semantic content [37]. While the context may be no more than the type of media that the data represent, such as text (characters) or a bitmap image (pixels), some contextual knowledge must be available in the system; otherwise, the digital inscriptions cannot be computationally processed and are not recognizable as data at all. However, and despite these remarks, it is important to stress that our view of data is consistent with much of extant literature [1, 5, 12, 23].

We view data as semantic material or a resource that inscribes external facts [4]. The defining attribute of data is thus their capacity to represent things or events other than themselves, which – we claim – stems from a structure that embeds knowledge of how each *data token* (datum) stands for something [37]. For instance, temperature can be recorded as data only if the measurement apparatus in use embeds the knowledge of what does it mean to measure temperature, and the data must retain a connection with such knowledge or lose their capacity to represent temperature. Devoid of such contextualizing structure, temperature records are nothing but meaningless numbers. We may thus ask: *What makes data possible?*

To seek answers to this question, we develop a perspective that problematizes a capacity to structure data. The perspective acknowledges the deeply sociotechnical nature of data [32], and allows to study data innovations as distinct from the broader but closely related category of digital innovations [22, 26, 40, 41]. We begin our discussion by drawing on computer science literature on semi-structured data [2, 28], an emerging stream of research on data-in-practice [24], and the theory of digital objects (e.g., [18, 25]) to conceptually unpack the idea of structured data. We then present eXtensible Business Reporting Language (XBRL) as a vignette to illustrate the ways in which data structures are enacted such that they make it possible to produce data that are useful for financial reporting. We show that rather than assuming that data 'has' (or has not) a structure [37], data are better viewed as made by a structure that gives digital inscriptions a capacity to represent specific contextual facts. Consequently, new ways to structure data for a particular context can unlock new ways to create value through the data, that is, phenomena we refer to as data innovations.

2. A Computational and Social View of Data

A standard view in the literature that serves most practical purposes well is that data are raw unorganized facts [1, 12] or invariances [23] from which information and, ultimately, knowledge can be extracted [see also 5, p. 109]. The term *unstructured data* (see Appendix A for definitions of key concepts) is generally used for any semantic content, whether as a separate file or records embedded in an executable code, whereas *structured data* are normalized records that reside in a database system subject to rigid and regular structure [2, 28]. The latter are accessed through a database engine that enforces a common schema – that is, a *data model* by which each individual data token is restricted to a set of attributes that adhere to the schema [7]. Structured data can encapsulate unstructured data such as fields for natural language content and bitmap image data, meaning that the difference between the two types of data can also depend on the granularity at which the matter is observed. Also, it may be illuminating to note that the underlying files that store the data accessed through a database lack much of the structure that the database engine imposes at the time of use.

Between the extremes of structured and unstructured data, computer science recognizes *semi-structured data*, characterized as "schema-less or self-describing" data [3]. This means that data are not accompanied by a robust type and structural description but an explicit structure is otherwise present in the data. What separates semi-structured data from unstructured data is that the former is structured in a manner that can be perceived by observing the data itself [2]. Relative to structured data, semi-structured data can be characterized by irregularity and instability of structure, an implicit and *a posteriori* schema (as opposed to one that is precisely specified *a priori*), that is, a 'sketchy' data structure that often hampers interoperability [2].

Some of the issues associated with lack of appropriate structure in data can be tackled *ex post*. Computer and data scientists have developed various techniques, for instance, to discover structures in unstructured text [28], extract structured data from web sites [7, 42], and to recognize patterns in images [13], to name a few examples of work that seek to recover a structure from seemingly unstructured data. There are tools to detect changes in data schemas over time and, thereby, tackle issues of rapidly evolving, unstable data structures [10]. Researchers also continue refining techniques to query and extract information from unstructured and semi-structured

data alike [11, 39]. All in all, while computer and data science can tell us how a structure can be imposed on or extracted from data, they do not explain how or why a particular way of structuring data renders them useful in an industry or organizational setting.

The problem of structuring data cannot be treated as a technical issue alone; instead, the data need to be understood as a human creation that is entangled with social practices and the institutional setting in which the data are used [23]. This means that speaking of raw data as a sort of *de facto* natural resource is misleading [20] as it tends to obscure organizational processes, innovations, and work involved in making data effective inside and between organizations [24]. As we have discussed above with respect to temperature, recording seemingly simple facts about, for instance, a company's financial results requires that we know a lot about local accounting laws, regulations and practices – knowledge that is embedded in how we structure the data [20, 37].

Jones [24], recently called for research on data-in-practice that frames data and their use in terms of two questions: *How data come to be? How data come to be used?* The former question refers to work, practices, and decisions that create, maintain, and replenish data sources; the latter question points to issues associated with how data are actually used in organizations. Rather than being an idle resource waiting to be accessed in a database, data are often ambiguous and performed to different ends according to the data-in-practice perspective. Accordingly, Gitelman [20, p. 7] notes that one "productive way to think about data is to ask how different disciplines conceive their objects, or, better, how disciplines and their objects are mutually conceived."

Extending these ruminations, we argue that there is a third important type of questions that data-in-practice research needs to engage: *What kinds of preconditions need to be present for data about a phenomenon to exist?* Also, studying this question in an empirical setting entails answering: *What structures make specific real-life data possible?*

3. Data as Digital Objects

Let us define data as digital objects that have a capacity to carry facts about the external world. The definition is largely consistent with the above-mentioned textbook view of data as "raw facts that describe a particular phenomenon" [21, p. 508], "a series of facts that have been obtained by observation or research and recorded" [9, p. 794], or "raw facts that can be processed into accurate and relevant information" [38, p. G-3].[1] However, scholars have recently called more careful attention to the 'factness' of data, arguing that data are not a sort of natural or foundational substance. Gitelman [20], Jones [24], and those taking the tack of Tuomi [37] make the point that data are human-made and bound up with specific practices and institutional settings. To study data from this perspective, we take a look at i) what are the constituent parts of data, and ii) how do these come together as data objects with a capacity to represent external facts.

We use a data token (datum) as a generic term for the constituent entities that make up data. An alpha-numeric character or a sequence of them is the most common type of token but by no means the only one – for instance, an encoded pixel in a bitmap image can be similarly seen as part of a larger data object. At the same time, not just any collection of alphanumeric characters or pixels counts as data: to bring data tokens together as data, something more than just the constituent elements is needed. We use the term *data object* to refer to a collection of data tokens that is present in social practice as a thing. Actors can identify the object in their ongoing practices, and the data object can become a resource, constraint, or otherwise involved in the practice. For instance, a data scientist who perceives a collection of alphanumeric characters as relevant data may be able to use a collection of tweets as a resource to build a sales forecasting model. Note that such everyday 'objectification' takes largely place by virtue of habit and routine that provide a social infrastructure for the smooth operation of organizational life.

The theory of digital objects defines objects as structured continuants [17, 18]. First, an object is an arrangement of other objects; that is, it has a structure which gives rise to emergent properties such as a capacity to represent facts. Second, the object endures at least for a period of time that allows actors to treat it as an object-in-practice. For instance, a web page object is an arrangement of text and images that lasts at least as long as the page is loaded into a web browser, allowing a user to assess the content of the page. An object's life span can range from very short, as in the case of an individual search engine results page, to theoretically infinite, as in the case when the page is archived for future reference [25].

Faulkner and Runde [18] call entities at the most rudimentary level of computation bitstrings. Bitstrings are series of binary distinctions encoded in a material medium. They are the link between physical things and the realm of computing, which

[1] The examples are from Jones [24].

provides a necessary material footing for digital objects such as data to exist. Furthermore, bitstrings are syntactic objects whose constitution is governed not by their physical attributes but by a language that specifies how parts may be arranged into higher-level objects. In any IT equipment there are many such languages embedded (e.g., character encodings) by which the equipment can automatically transform rudimentary binary distinctions into digits in a binary number system, from the series of digits into numbers, and from numbers to alphanumeric characters that can then constitute many other types of objects including data objects. We may call a language governing the constitution of digital objects a code whenever it is embedded as a standard part of the computational equipment.

However, neither bitstrings nor alphanumeric characters, or any combination thereof, are 'raw data' in the sense of unmediated or plain facts. It should be clear from what has been said above that syntactic objects including data are couched in a natural or formal language (or code) that is always a human creation [18] including numerous choices that empower and limit the expression of facts by the data. For instance, the original ASCII character encodings were limited to 26 letters characters in the English alphabet and could not express Scandinavian letters such as 'ä' or 'ö'. The languages and codes involved in the construction of a data object define the ways in which the object can represent external facts.

A database schema or a data model that governs how structured data capture facts from a particular domain is another example of such a language, but – and this is central to our argument – no data can exist without being couched often in multiple interwoven languages and codes that give them the power to represent external facts. For instance, natural language text is often considered unstructured data from the perspective of analytics, yet it must at minimum i) follow the rules of English or some other human language and ii) adhere to a character encoding if it is not to be mere gibberish. To reiterate, we often refer to such languages as codes if they are embedded into the computation equipment itself, which also tends to make them somewhat invisible yet without them information systems could not operate.

Implicit in much of the foregoing is the idea that digital objects are layered entities [17, 18]. We have distinguished among bitstrings, binary digits, numbers, alphanumeric characters and pixels as progressively more aggregate entities; however, the layering applies equally to much more complex objects such as documents of all kinds. For instance, a PDF document is a complex object based on the PostScript language in addition to the rudimentary entities listed above. The document can further act as a bearer for other types of syntactic entities, such as a company annual report that must additionally conform to the rules and regulations of the respective accounting domain. The composition of the lower-level object (PDF document) largely determines which kinds of operations can be performed on the higher-level object (annual report). For example, it is possible to copy and paste text from the annual report rendered as a PDF document, while a paper printout affords a different set of operations on the same report.

Physical things such as a hard copy of company annual report gain 'objecthood' fairly easily due to the relative stability afforded by a material bearer and shared conventions formed around the physical rendition of the object. Note that by rendition (or rendering) we refer to an instance of a syntactic object that is borne by a specific medium. Others, especially those with non-material bearers, can be much more ambivalent as objects. Lacking spatial attributes, digital objects are also often distributed so that it can be difficult to say where one object ends and another begins as their parts may be brought together as objects only in practice [15, 25]. Take, for instance, a database engine that creates structured datasets in response to specific queries instead of storing dataset objects themselves. The user or another computational process requests the rendering of the data for a particular purpose; in this sense, the data tokens do not 'have' a structure but are embedded in one that allows making sense of them in real time.

Another important aspect of data is that data are always about something and data for something. This is to say that data are defined as technological objects by their capacity to represent things for one or more (analytical) purposes. This derives from a generic assertion that technology is a means to an end and that, to be recognized as such, a technological object must express a distinct instrumental character [16, 29, 35]. More specifically, the identity of a technological object results from a collectively assigned function [17] that, in the case of data, hinges on the capacity of data objects to represent relevant external facts. This capacity, in turn, is based on a structure or a capacity to impose a structure on alphanumeric characters or other types of data tokens in such a way that they make sense in a given context. This is associated with how cognitive science describes human information processing in drawing a distinction between internal embodied and external declarative schemata that endow us with the capacity

to make sense by categorizing stimuli. Internal schemata involve beliefs, ideologies, and language [27], whereas external schemata are artifacts, social rituals, practices, and other embodiments of collectively held conventions. Schemata are mental shortcuts required for organizing and processing incoming information and perceptions in light of existing knowledge structures and processes related to contextually relevant entities.

Now that we have described the constituent parts of data and how these come together as data objects that have a capacity to represent relevant facts, we move to present the case of XBRL as a concrete example showing how a new type of data (for financial reporting) became possible.

4. Structured Data in Financial Reporting – the Case of XBRL

eXtensible Business reporting Language (XBRL) is a popular domain-specific language for storing financial data in structured format and making the data interoperable between organizations [14]. XBRL is based on eXtensible Markup Language (XML), and it uses XML syntax and related XML technologies such as XML Schema, XLink, XPath, and namespaces. The most common use cases for the language are found in government-mandated reporting of aggregated data such as financial statements, tax reports, and the provision of other statistics in a machine-readable form. Also, XBRL provides tools for transactional reporting. Overall, the development and adoption of the language offers a good illustration of what we call a data innovation: a new capacity to structure data in such a manner that they can create analytical insights in a specific context.

4.1. The Evolution of XBRL

XBRL originated in July 1998 when Charles Hoffman approached the American Institute of Certified Public Accountants (AICPA) with the idea of describing financial statements and audit schedules via XML. At that time, many firms had just started to utilize the internet for financial reporting by presenting key indicators and other information on their web sites. The most important document in this regard is the annual financial statement, which provides key information to firm's shareholders and other stakeholders.

The initial idea was to disseminate the information contained in the annual financial statements more efficiently. Hoffman was invited to brief the AICPA's High Tech Task Force on XML in September 1998, and his proposal eventually led to the development of a prototype set of financial statements using XML together with a business plan for the use of XML in financial reporting in the US. The original plan, prepared by a group of certified public accountants including Hoffman (an independent CPA), Wayne Harding (with Great Plains), Eric Cohen (for Cohen Computer Consulting), and Louis Matherne (the AICPA's Director of IT), presented a business case and roadmap for XML-based financial reporting, which contributed to the formation of a formal steering committee focused on development of an XML-based financial reporting language. The committee was joined in August 1999 by large auditing firms such as KPMG and Ernst & Young that sensed the potential for the language to have a disruptive effect on the auditing profession, and by major technology firms such as Microsoft recognizing the business opportunity in XML-based infrastructure and data transmission. The prototype reports were completed in October 1999, when the financial statements of ten companies were converted into XML. The committee became officially the XBRL steering committee in April 2000, lending further credibility and institutional support to the development of XBRL in the accounting domain. [30]

Over the last two decades, XBRL has grown into a globally accepted language for expressing financial data. In the US, the accounting scandals of the early 2000s and subsequent legislation requiring more prudent and transparent financial reporting and auditing have significantly fueled the growth of XBRL. In Europe, a recent EU transparency directive has paved the way for enforcing publicly listed companies' use of XBRL in their financial reporting through national legislation. These developments have been made possible and further supported by the availability of several XBRL-compatible financial, tax, and statistics reporting software packages developed by different software vendors.

4.2. XBRL as a Capacity to Structure Data

Financial information captured in XBRL must adhere to a taxonomy governed by XBRL International and its local consortia ("jurisdictions"). Facts are stored in an XBRL instance document that structures the information by means of descriptive and structural metadata. The instance documents are machine readable; that is, an XBRL-compatible software can read the data structure by referring to the metadata provided. The instance document can be stored as a standalone file in a file system or, for

instance, embedded in other documents, such as HTML pages.

Descriptive metadata are defined by a taxonomy schema that articulates the semantic meaning of data tokens that the instance document is allowed to carry. The entities to be reported in, for example, a firm's annual report, are connected to the corresponding data elements in the XBRL taxonomy schema through a process called tagging. Data tokens are created by giving recorded values distinct definitions through tagging and, by implication, a capacity to represent external facts; that is to say, it is the structure to which a certain set of recorded values belongs through tagging that makes them 'data' in the accounting context. In addition to the basic definition of a data token, an XBRL taxonomy schema entity such as "Deferred Tax Assets, Net" often stipulates further descriptive metadata such as currency, periodicity, and credit/debit status for the token.

Structural metadata are provided through so-called XBRL linkbases that articulate valid relationships between data tokens within an instance document, and between data tokens and external resources. There are five main types of linkbases: a label linkbase provides human-readable descriptive strings for data tokens, a reference linkbase connects data tokens to authoritative literature such as accounting laws, a calculation linkbase associates data tokens with each other so that values can be checked for consistency, a definition linkbase expresses the relations between data tokens, and a presentation linkbase facilitates the rendering and visualization of the data.

The two parts of the XBRL taxonomy (the schema and linkbases) need to be localized to address the fact that accounting laws and practices differ from a country to another. Hence, governing the XBRL schema and associated linkbases is far from a technological matter alone and requires deep understanding of the national regulatory environment and accounting practices. This further highlights the social nature of data structure: the layers of descriptive and structural metadata, and their enactment in accordance with local accounting laws and regulations form a sociotechnical system in which the technical components of descriptive metadata (the schema) and structural metadata (the linkbases) are constituted in interaction with social practices and institutions aimed at maintaining transparency and control of financial data and bookkeeping.

4.3. Extending and Adapting XBRL

As its name and roots in XML suggest, XBRL is extensible. Firms can extend the XBRL taxonomy by adding their own entities to it. The extensibility allows making XBRL data more expressive internally, but it can also weaken the comparability of instance documents across firms. As a result, a technique called anchoring has been mandated for recent XBRL deployments: whenever a firm chooses to extend the national XBRL taxonomy with a new firm-specific entity, this needs to be mapped (that is, 'anchored') to the nearest available entity in the national taxonomy.

XBRL instance documents are constructed to be machine-readable, but humans often need to read the content of documents for auditing and other purposes. While the XML foundations of XBRL make the language human-readable to some extent, inline XBRL, or iXBRL, was developed by the XBRL community to facilitate the rendering of data contained in XBRL instance documents in a way that is easy to understand for humans. In an iXBRL instance document, XBRL data is embedded into an HTML document.

Finally, the data within XBRL instance documents can be structured to the degree desired. For instance, a firm may choose to tag the body of its annual financial statements (i.e., the income statement and balance sheet) in detail, thus converting these into several, highly granular and machine-readable data tokens, while opting to use only block tagging for the notes to the financial statements. Block tagging marks each section as a whole (e.g., identifying the CEO's letter and the auditor's report), whereas any financial details inside the blocks are not part of the data structure and therefore not computationally accessible.

4.4. Changing Auditing Practices

Law typically requires that the financial statements of most publicly listed companies and some private ones (typically companies above a certain size threshold, which depends on local legislation on statutory reporting to the government) are audited by an external auditor. The audit provides assurance to shareholders and other stakeholders (e.g., government authorities and business analysts following the company) by verifying that the statements record a good and fair portrayal of the company's financial situation. Moving over to use an XBRL instance document instead of a paper or PDF document as the audited object has significant ramifications for auditing. Although XBRL grants

auditors new opportunities to use more powerful tools to access company financial details subject to auditing, it also imposes additional competency requirements. Working with XBRL data schemas and linkbases requires the auditor to possess at least rudimentary IT skills.

In conjunction with changes in the nature of financial data, a debate has emerged within the audit community about what should be the object for auditing and how to demarcate the boundaries of an audit when the data are provided as an XBRL instance document that is inherently distributed in nature [25].[2] An important aspect of the discussion is the verification of the tagging procedure explained above and, consequently, the definition of a legal document in the context of financial reporting. Auditing bodies have started to debate whether the audit of a firm's financial statement should include validation and a stance on whether the data tokens in the firm's financial systems are correctly connected to the XBRL taxonomy schema. Furthermore, opinions differ on whether a statement offering such assurance should be part of the formal audit report or, instead, contained in a separate report with a different legal status. As for the nature of the legal document audited, there are several views on what kind of object ought to be archived as the official financial statement: a machine-readable XBRL instance document, a physically signed hard copy, a digitally signed version (possibly in PDF format), or something else.

5. Discussion and Conclusions

We have problematized the notion of a clear-cut distinction between unstructured and structured data and presented an argument that data do not 'have' a structure but a structure or a capacity to structure data at the time of use is what makes data. It follows from this that there are no literally unstructured data. Understanding how digital data gain structure is highly relevant for theory and practice alike at a time when data appear to be increasingly driving value creation across industry boundaries.

A general observation emerging from our theoretical analysis and XBRL vignette is that the production of data is an inherently sociotechnical process [32]. On the one hand, the analysis shows that data are entangled with the details of technical implementation so tightly that one cannot fully understand them in isolation from the systems that render the data objects. On the other hand, what makes certain digital objects 'data' is their capacity to represent external facts. The semantic or sensemaking potential of data is conferred by a structure that establishes a connection between the data tokens and a domain of human activity and hence turns digital inscriptions into data about something.

We make several important observations regarding how data gain a structure. The first is that the structure is always relative to a purpose that makes sense in a specific context. In this sense, while our example of XBRL represents a domain-specific language, the key observations we draw from the case are not limited to domain-specific languages. The purpose may be a seemingly simple matter of representing alphanumeric characters in an IT equipment (character encodings) or a complex societal matter such as representing financial information in a manner consistent with the local regulatory environment (XBRL). In this sense, all data must be structured in some way, since without structure digital inscriptions cannot represent facts and are thus not recognizable as data. What is usually meant by 'structured data' is data that are structured by recourse to an external language, schema, or data model, whereas data that are structured only by codes internal to the IT equipment are often seen as unstructured data.

We further note that individual data tokens such as characters and words in a prose or the pixels of a bitmap image can be part of multiple structures that are enacted by, for instance, natural language processing or pattern recognition technologies; at the same time, robustly structured data such as inline XBRL documents can be embedded in loosely structured data such as web pages. Finally, the same digital inscriptions may be structured as data to different degree with respect to different purposes.

Against this background, XBRL is an example of what we suggest to call data innovations, that is, a new valuable way to render phenomena as data that

[2] In connection with implementation of the European Union's Transparency Directive (2013/50/EU), the European Securities and Markets Authority (ESMA) mandates publicly listed companies in the EU area to prepare XBRL-tagged financial statements from 2020 onward. Since ESMA's announcement of the mandatory reporting program, various auditing bodies have been engaged in far-reaching debate on the requisite extent of auditing. To support our discussion, we provide the reader with links to opinions expressed by the Committee of European Auditing Oversight Bodies (https://ec.europa.eu/info/sites/info/files/business_economy_euro/banking_and_finance/documents/191128-ceaob-guidelines-auditors-involvement-financial-statements_en.pdf), Accountancy Europe (https://www.accountancyeurope.eu/wp-content/uploads/191217-ESEF-assurance-paper-FINAL_update_2.pdf), and a local audit community (https://www.suomentilintarkastajat.fi/content/download/34450/1052269/version/1/file/Listayhti%C3%B6n+ESEF-tilinp%C3%A4%C3%A4t%C3%B6ksen+varmentaminen+ST+suositus+2_2020.pdf, in Finnish).

can be processed computationally. The case vignette reveals structures (descriptive and structural metadata, governance processes, etc.) that allow recording financial information in a machine-readable and interoperable format that has enabled significant advances in accounting and auditing professions. While closely connected with the more general category of digital innovations, data innovations are a distinct category that we believe is worthy of consideration in its own right. For instance, modularity – which is a core principle and enabler of digital innovations – gains a different meaning in the context of data innovations [4]. Modularity entails breaking a complex system into simple components connected by clearly defined interfaces to enable complex functionality [6, 33, 34]. Modularization makes system components internally manageable and allows one to (re)combine them in multiple ways, which typically enable faster system adaptation and innovation. While there are obvious parallels to modularization in the making of structured data, the latter is driven by prospects to create meaningful rather than functionally complex combinations, which cannot be understood by recourse to the standard logic of modularity alone.

Finally, despite major practical differences between what is currently known as unstructured and structured data as economic resources, management scholars have until now devoted little attention to the distinction that we have attempted to deconstruct in this paper. Digital data are more and more often the raw material from which things are made, resulting in what Baskerville et al. [8] call ontological reversal. The digital versions of things (such as financial statements in the form of XBRL instance documents) become primary institutional objects forcing human practices to adapt accordingly to the new material form and behavior of objects.

To conclude, we argue that the IS discipline could evolve to provide, in a sense, the material science of digital economy, reflecting the important recognition in this and other recent papers [4, 20, 24] that data are a more complex matter than previously thought. For instance, approaches such as the theory of digital objects [18] and digital operations [31] can tease apart the nature of data as non-material entities defined by semantic capacity to represent external entities and in so doing pay due attention to the social, technical, and economic aspects of data. Conceptualizing data innovations creates avenues to answering the question of what makes data possible – which we argue should be the third dimension in the study of data-in-practice [24].

6. Appendix A: Key concepts and definitions

Concept(s)	Definition	Example from accounting
Unstructured data	A mishmash of semantic entities that can differ from an observation to another; it may not always be clear what constitutes an individual observation	Data residing in a note written with a text editor to be refined into a receipt to be booked into an accounting information system
Semi-structured data	Data organized using an irregular or unstable data structure which hampers the usability and interoperability of the data	Data residing in an electronic sales invoice adhering to a proprietary XML-format that needs to be converted to the XML-format of electronic purchase invoices
Structured data	Data residing, for instance, in a database under a rigid and regular structure with well-defined fields that correspond to distinct variables	Company's financial statement stored in a standardized, taxonomy-compliant XBRL instance document
Bitstring	Series of binary distinctions encoded into a material medium	Magnetic marks on a hard disk platter
Data token, raw unorganized facts, invariances	Data token refers to the most granular element of data; also called invariances as they remain unchanged when a specific transformation is applied	”Deferred Tax Assets, Net” in an XBRL instance document containing a company's financial statements
Data object	Aggregated or computed entity made out of data tokens	Key financial figure computed using data tokens such as return on capital employed
Metadata	Data that provide information about other data	Metadata in an XBRL instance document (e.g., currency, periodicity, and credit/debit status of Deferred Tax Assets, Net)
Data model or schema	Definition of the organization of data; articulates allowed data tokens and their attributes, and specifies the possible relationships between them	XBRL taxonomy (e.g., US GAAP XBRL taxonomy for financial statements)

Data source	A location from where the data being used originates	Relational database (e.g., the EDGAR repository for US GAAP XBRL financial statements)

7. References

[1] Abbasi, A., S. Sarker, and R.H.L. Chiang, "Big data research in information systems: Toward an inclusive research agenda", *Journal of the Association for Information Systems 17*(2), 2016, pp. i – xxxii.

[2] Abiteboul, S., "Querying semi-structured data", *International Conference on Database Theory*, (1997), 1–18.

[3] Abiteboul, S., P. Buneman, and D. Suciu, *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann Publishers, San Francisco, CA, US, 2000.

[4] Alaimo, C., J. Kallinikos, and A. Aaltonen, "Data and Value", In S. Nambisan, K. Lyytinen and Y. Yoo, eds., *Handbook of Digital Innovation*. Edward Elgar Publishing, 2020.

[5] Alavi, M., and D. Leidner, "Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues", *MIS quarterly 25*(1), 2001, pp. 107–136.

[6] Alexander, C., *Notes on Synthesis of Form*, Harvard University Press, Cambridge, MA, 1964.

[7] Arasu, A., and H. Garcia-Molina, "Extracting Structured Data from Web Pages", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (2003).

[8] Baskerville, R.L., M.D. Myers, and Y. Yoo, "Digital First: The Ontological Reversal and New Challenges for Information Systems Research", *MIS Quarterly 44*(2), 2020, pp. 509–523.

[9] Bocij, P., A. Greasley, and S. Hickie, *Business information systems: Technology, development and management*, Pearson Education, Harlow, England, 2008.

[10] Chawathe, S.S., and H. Garcia-Molina, "Meaningful Change Detection in Structured Data", *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 1997.

[11] Chen, Y., W. Wang, Z. Liu, and X. Lin, "Keyword search on structured and semi-structured data", *SIGMOD-PODS'09 - Proceedings of the International Conference on Management of Data and 28th Symposium on Principles of Database Systems*, (2009).

[12] Davenport, T.H., and L. Prusak, *Working knowledge: How organizations manage what they know*, Harvard Business School Press, Boston, MA, 1998.

[13] Egmont-Petersen, M., D. De Ridder, and H. Handels, "Image processing with neural networks- A review", *Pattern Recognition 35*(10), 2002, pp. 2279–2301.

[14] Eierle, B., H. Ojala, and E. Penttinen, "XBRL to enhance external financial reporting: Should we implement or not? Case Company X", *Journal of Accounting Education 32*(2), 2014, pp. 160–170.

[15] Ekbia, H.R., "Digital artifacts as quasi-objects: Qualification, mediation, and materiality", *Journal of the American Society for Information Science and Technology 60*(12), 2009, pp. 2554–2566.

[16] Faulkner, P., C. Lawson, and J. Runde, "Theorising technology", *Cambridge Journal of Economics 34*(1), 2010, pp. 1–16.

[17] Faulkner, P., and J. Runde, "Technological objects, social positions, and the transformational model of social activity", *MIS Quarterly 37*(3), 2013, pp. 803–818.

[18] Faulkner, P., and J. Runde, "Theorizing the digital object", *MIS Quarterly 43*(4), 2019, pp. 1278–1302.

[19] Forbes, "What Will We Do When The World's Data Hits 163 Zettabytes In 2025?", *Forbes*, 2017.

[20] Gitelman, L., *Raw data is an oxymoron*, MIT Press, Cambridge, MA, 2013.

[21] Haag, S., and M. Cummings, *Management information systems for the information age*, McGraw-Hill, Inc., New York, NY, 2013.

[22] Henfridsson, O., J. Nandhakumar, H. Scarbrough, and N. Panourgias, "Recombination in the open-ended value landscape of digital innovation", *Information and Organization 28*(2), 2018, pp. 89–100.

[23] Hirschheim, R., H.K. Klein, and K. Lyytinen, *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations*, Cambridge University Press, Cambridge, UK, 1995.

[24] Jones, M., "What we talk about when we talk about (big) data", *Journal of Strategic Information*

Systems 28(1), 2019, pp. 3–16.

[25] Kallinikos, J., A. Aaltonen, and A. Marton, "The Ambivalent Ontology of Digital Artifacts", *MIS Quarterly 37*(2), 2013, pp. 357–370.

[26] Kohli, R., and N.P. Melville, "Digital innovation: A review and synthesis", *Information Systems Journal 29*(1), 2019, pp. 200–223.

[27] Larson, D.W., "The Role of Belief Systems and Schemas in Foreign Policy Decision-Making", *Political Psychology 15*(1), 1994, pp. 17–33.

[28] McCallum, A., "Information Extraction: Distilling Structured Data from Unstructured Text", *ACM Queue November*, 2005.

[29] Orlikowski, W.J., and C.S. Iacono, "Desperately Seeking the 'IT' in IT Research - A Call to Theorizing the IT Artifact", *Information Systems Research 12*(2), 2001, pp. 121–134.

[30] Roohani, S., "Section Six: What is the History of XBRL?", *XBRL Education*, 2008. http://xbrleducation.com/edu/history.htm

[31] Salovaara, A., K. Lyytinen, and E. Penttinen, "High reliability in digital organizing: Mindlessness, the frame problem, and digital operations", *MIS Quarterly 43*(2), 2019, pp. 555–578.

[32] Sarker, S., S. Chatterjee, X. Xiao, and A. Elbanna, "The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance", *MIS Quarterly 43*(3), 2019, pp. 695–719.

[33] Schilling, M.A., "Toward a general modular systems theory and its application to interfirm product modularity", *Academy of Management Review 25*(2), 2000, pp. 312–334.

[34] Simon, H.A., "The Architecture of Complexity", *Proceedings of the American Philosophical Society 106*(6), 1962, pp. 467–482.

[35] Stein, M.K., S. Newell, E.L. Wagner, and R.D. Galliers, "Coping with information technology: Mixed emotions, vacillation, and nonconforming use patterns", *MIS Quarterly 39*(2), 2015, pp. 367–392.

[36] The Economist, "The world's most valuable resource is no longer oil, but data", *The Economist*, 2017.

[37] Tuomi, I., "Data Is More Than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory", *Journal of Management Information Systems 16*(3), 1999, pp. 103–117.

[38] Turban, E., D.E. Leidner, E. McLean, J. Wetherbe, and C. Cheung, *Information Technology for Management: Transforming Organizations in the Digital Economy*, Wiley, Hoboken, NY, 2006.

[39] Yao, X., and B. Van Durme, "Information extraction over structured data: Question answering with freebase", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, (2014).

[40] Yoo, Y., R.J. Boland, K. Lyytinen, and A. Majchrzak, "Organizing for innovation in the digitized world", *Organization Science 23*(5), 2012, pp. 1398–1408.

[41] Yoo, Y., O. Henfridsson, and K. Lyytinen, "The new organizing logic of digital innovation: an agenda for information systems research", *Information Systems Research 21*(5), 2010, pp. 724–735.

[42] Zhai, Y., and B. Liu, "Structured data extraction from the web based on partial tree alignment", *IEEE Transactions on Knowledge and Data Engineering 18*(12), 2006, pp. 1614–1628.