Editorial: Managing unstructured information

Luke Tredinnick & Claire Laybats

When does more organisation lead to worse outcomes? Information and Knowledge Management were established on the principle that well-ordered collections confer significant benefits to individuals and organisations. A well-ordered collection makes it easier to identify the right resources, and easier to ensure that you are fully exploiting the entire information collection. It might therefore be assumed that it is always better to invest the time and effort into creating well-organised collections, particularly where the potential benefits outweigh the potential costs. But this is not always the case. Sometimes more highly ordered collections lead to worse outcomes. This is because highly-structured data-sets and information collections are more susceptible to various kinds of error and decay. Defining data in great detail allows more control but can also result in greater data drift and corruption. The more structure we impose the more ways there are for things to fall out of order.

This editorial explores the role of unstructured and semi-structured information in the workplace, and how to maximise their benefits. Unstructured information refers to data that does not conform to predefined models or schema, making it more flexible but also more challenging to manage. This type of information often consists of large volumes of text, images or other multimedia content making it rich in insights but difficult to search and organize. Typical examples may include reports, presentations, documents, notes, and insights. Semi-structured information on the other hand is a type of data that while it does not conform to a rigid schema does still contains tags or markers that separate semantic elements and enforce hierarchies. This kind of data lies between structured and unstructured data, combining some of the advantages of both. Examples include email, web-pages, social media posts, some documents, and some multimedia content.

The vast majority of information created and used within organisations consist of unstructured or semi-structured information and data. In 2023 a report by IDC highlighted that 90% of data generated by organizations was unstructured, accounting for over 57,000 exabytes of data globally (Muscolino et al, 2023). Much of this information is critical to organisations. It exists in client communications, supply chain management, marketing content, employee appraisals, in design and engineering documents, in opinions and expert advice and in many other areas of organisational activity. Unstructured and semi-structured information comprises the vast majority of records of human communications, including emails, minutes, meeting notes, presentations, social media content, transcripts, recordings, and videos. Organisations invest enormous resources and effort in creating this information; it is the material outcome of most of the work that is done. Yet unstructured and semi-structured information and data is often under-utilised within organizations, sitting passively in disparate information systems, unheralded in content management systems, and isolated in personal information silos. Those ad hoc messy collections are of critical importance to the organization, distilling the professional expertise and intellectual labour of its workers. Yet almost half of the information created within organizations is barely used after first creation and its initial use (Muscolino et al, 2023). Capturing these insights and transforming them into a more accessible resources could revolutionize the ways organisations function. Effectively managing unstructured and semi-structured information is therefore an emerging issue in commercial information and knowledge management.

Unfortunately unstructured information presents a number of significant challenges to its effective organisation and use. Tanwar et al write that “unstructured data adds to the complicacy of analytics process since mostly the machines require a structural organization of data for processing and analysis” (2015). It tends to be rich in meaning, significance, and

nuance, and therefore requires more interpretation in order to understand its relevance to any given task. This richness is partly because unstructured and semi-structure information tends to be highly context-dependent. Understanding its full significance depends upon understanding the context in which it was produced and the uses to which it is put. The relevance of an existing report to a given problem for example is in part determined not only by its content, but also the reasons why that report was produced, the questions that it sought to resolve, the audience for which it was originally intended, and the ways in which it was originally used. Because of this context-dependent nature, unstructured and semi-structure information is often most valuable when connected to other relevant information, records and insights. It is in general less discrete, and more interconnected, and tapping its potential often relies getting it into the hands of the right people. The insights from an individual email for example may be easier to identify when connected to other emails in a chain, or to other documents from the same source, to other information that pertains to the same topic, and when it is interpreted by someone involved in its production. This means that the significance of unstructured and semi-structured information is not readily apparent, and often emerges from the ways in which it combines in a complex network on meanings and semantic associations. Finally semi-structured and in particular unstructured information is highly heterogeneous, coming in many different forms.

These characteristics make it more difficult to identify and categorise the types of unstructured and semi-structure information within an organisation, ensure data quality, implement robust security and ensure compliance with data protection regulations, provide resource descriptions, to classify, index or organise collections of unstructured information and to identify what information are likely to be of future significance and which are not. While much of the work in managing structured data resources such as databases and information systems is done in advance ensuring a well organised and well-managed resourced that can be well-understood in principle, unstructured information has traditionally required much of that cognitive work to be done on-demand in relation to specific queries. This makes unstructured information challenging and resource-intensive to work with. Far easier to consult the insights from an enterprise management system than to consult the relatively richer data from innumerable business critical reports that may have been prepared for specific purposes in other contexts but whose history and significance has been long forgotten.

It is partly these kinds of issues that the World Wide Web was originally intended to address. The World Wide Web was created as a means of managing information at the European Centre for Nuclear Research (CERN) in the mid-1980s (Berners-Lee, 1999). CERN presented a specific challenge for information management; because the facility housed a continually shifting population researchers. In the original proposal Berners-Lee described CERN itself as ‘a multiply connected ‘web’ whose interconnections evolve with time’ (1989), and later reflected on CERN’s ‘weblike structure’ (1999:10) and its ‘complexity’ (1999: 150). The web itself was intended as a more organic means to capture these shifting relationships and connections, in part because it mirrored the situation that it set-out to address. Berners-Lee later reflected:

> I had seen numerous developers arrive at CERN to tout systems that ‘helped’ people organise information. […] I saw one protagonist after another shot down in flames by indignant researchers because the developers were forcing them to reorganise their work to fit the system (1999: 17)

The Web reflected the needs of its users by growing organically through the semantic and associative connections made by users themselves. Nevertheless there are in fact two kinds of structure that emerges in the original Web design: HTML which structures documents with

semantically meaningful labels, and hyperlinks which create an organic structure between and within individual documents on the basis of semantic and associative connections.

Nevertheless the World Wide Web although revolutionary in its own terms did not fully address the problems of managing unstructured information. Analytical classifications were re-imposed through search-engine indexes (cf. Hess, 2008) and through hierarchical navigation schemes; the way in which the Web implemented its hyperlinks encouraged this imposition of hierarchical structure. As such hyperlinks became less important to resource discovery; while of course hyperlinks exploit semantic and associative connections in order to overlay a semantic and associative framework on the mass of information one, in fact over time our dominant means of engaging and seeking-out with web-content has shifted from browsing to search. The structure of hyperlinks provides and means to refine and narrow our interest within a specific results set, but search is the predominant means by which we identify a results set from the vast array of web content in the first place. And this highlights how full-text indexing and search have become the predominant means by which we manage unstructured and semi-structure information resources.

As a consequence our approach to managing unstructured information resources has largely fallen back on full-text indexing and search, with various approaches to semantic indexing intended to improve discoverability. Document and content management systems while promising to free-up productivity and improve information management have often acted to inhibit information sharing and collaboration by adding additional barriers. A permanent record that is difficult to locate as part of a large collection is often less practically useful in any given process than a temporary record that can be kept at hand and then forgotten. Martin White's (2023) paper on Workarounds published in *Business Information Review* last year highlights the common ways in which employees avoid complex workflow processes including relying on personal relationships and developing their own file systems and information silos. Solutions that drive collaboration can sometimes achieve the opposite.

However that limitation in managing unstructured and semi-structured information is now changing, largely because a new generation of AI tools enable computers to better understand the context of information and automate the kind of contextual connections that drive business productivity. The integration of document and content management with other business processes such as business insights, communication, and productivity tools means that unstructured information can be fed back directly into the productivity cycle. And because AI solutions are being integrated into common productivity packages, barriers to adoption are in many cases effectively eliminated.

There are a number of ways in which AI tools can aid organizations in managing unstructured information is an deriving business insights and benefits from that data. Natural Language Processing (NLP) systems and Large Language Models use AI techniques to understand, interpret, and manipulate language, important in processing and understanding text-heavy unstructured data. This may include for example information extraction and summarizing tools, such as providing concise overviews of complex documents, performing sentiment analysis, and extracting information for incorporation into more traditional analytical tools. NLP tools are increasingly embedded in productivity software to allow summaries of complex documents to be rapidly produced. A good example of this is Microsoft Meeting Insights which recommends contextual information. Tools like IBM Watson and Google's Natural Language API allow businesses to sift through large volumes of text to extract key information, provide concise summaries, and assess sentiment in social media posts or customer feedback. Similarly, Machine Learning algorithms learn from data patterns and apply this knowledge to categorize

unstructured data, detect anomalies, and make informed predictions, thereby enhancing data management and decision-making processes.

AI's role in managing unstructured information extends to data integration, automation, advanced analytics, and compliance. Tools like Tableau and Power BI help integrate unstructured data from various sources into cohesive formats, creating structured representations like knowledge graphs that connect entities and their relationships. This enriched data becomes more useful for analysis and decision-making. Automation platforms such as UiPath and Automation Anywhere streamline routine tasks, such as processing invoices, resumes, and contracts, as well as managing and prioritizing emails, significantly boosting efficiency and reducing manual effort. Advanced analytics provided by AI tools like SAS Analytics offer deeper insights into unstructured data, enabling predictive analytics to forecast trends and behaviors based on historical data and contextual analysis to understand the environment in which data is generated. In the domain of compliance and risk management, AI solutions like IBM Guardium and Varonis help identify and protect sensitive information within unstructured data, ensuring adherence to data protection regulations and assessing potential risks and vulnerabilities. Ultimately, AI's ability to process and manage unstructured information transforms how organizations handle and leverage vast amounts of data, leading to better decision-making, improved efficiency, and new opportunities for innovation.

September Business Information Review

The September issue of *Business Information Review* has a well-ordered focus on data. Data has been called “the new oil” (Brownlow et al, 2015); data driven businesses have become central to the ways in which we organised commercial operations in the twenty-first century. Sadowski for example has argued that “companies are clamouring to collect data – as much as they can, wherever they can” (2019). And yet as we have previously explore in the journal, data can also be dangerous (Tredinnick & Laybats, 2023). It invites misinterpretation and misunderstanding not least because it too glibly connotes rational, analytical precision, and often too neatly eradicates the social contexts in which productivity and organisational effectiveness flourish. The rise of data is therefore a key issue in commercial information management, and information science with its attention to the very human contexts in which information is created, disseminated and used an important counterweight to the statistical precision of data science.

Our first research paper in the September issue is entitled “The Role of Data Governance in a High-Level Approach of Data Migration to Open Data”. The paper explores the critical role of data governance in high-level data migration processes, stressing the importance of addressing data quality issues beyond technical aspects. It outlines the roles and responsibilities in the data migration process, key activities involved, stakeholders, and stages of data migration, emphasizing the critical role of data governance in ensuring project success by addressing data quality issues, establishing policies, and defining roles and responsibilities throughout the migration process. Our second paper is an opinion article, exploring the transformative impact of Artificial Intelligence and Machine Learning on libraries. Entitled “From Big Data to Intelligent Libraries: Leveraging Analytics for Enhanced User Experiences” the paper addresses artificial intelligence applications in cataloguing, search, and personalised recommendations, highlighting benefits and cahllenges of emergent technology. The paper argues that “The concept of intelligent libraries represents a significant evolution in

the role and function of libraries. It is not just about integrating new technologies, but about reimagining libraries as dynamic, adaptive, and user-centric information and learning hubs."

Our third paper is a research article from Philip Siaw Kissa of the University of Education Winneba, Ghana. Entitled "Examine the influence of collaborative business culture and data-driven analytic capability on business innovation: Moderation role of managerial capability", the paper discusses the relationships between data-driven analytics capability, collaborative business culture, and business innovation, highlighting the mediating effect of collaborative business culture. In addition it emphasizes the importance of balancing investments in data-driven analytics and managerial capability to enhance collaboration and drive innovation within organizations.

Returning contributor Mostafa Sayyadi has authored our fourth paper, a professional article exploring Machine Learning and Predictive Analytics. "How to Improve Data Quality to Empower Business Decision-Making Process and Business Strategy Agility in the AI Age" discusses the importance of data quality in empowering business decision-making and strategy agility in the AI age. It highlights the use of machine learning and predictive analytics to enhance data quality, management, and model interpretation. In addition it emphasizes the implementation of solutions like Explainable AI, cloud computing, and distributed systems to address data quality challenges and improve decision-making processes for sustainable growth.

The role of AI is also the theme of our next research article, "Exploring Ethical Considerations in AI-driven Cataloguing and Classification with ChatGPT". It explores the ethical considerations of using AI-driven cataloguing and classification systems, focusing on the use of ChatGPT in library sciences. The paper highlights privacy, accountability, transparency, and bias issues associated with these technologies, emphasizing the need for evolving ethical guidelines to align with technological advancements. Our sixth paper is an Opinion Article, entitled "The Use of Big Data in the Management of Library Resources". It explores the use of big data in the management of library resources in university libraries.

Our final paper is an Initiatives Article, and a change of focus for the issue: "The Green Library Advocacy: Need to Engage the International Communities in a Climate Change Action". Action on climate change is an increasingly important part of information and knowledge management, as highlighted in our recent editorial on *Decarbonising Information Work* (Tredinncik & Laybats, 2024). Ajani et al explore the pivotal role of libraries in global initiatives to combat climate change and emphasize the significance of international cooperation in advocating for sustainable practices.

BIR Best paper prize

We're delighted to announce that the Business Information Review Best Paper Prize for 2023 has been awarded to Martin White for the paper "Workarounds and shadow IT – balancing innovation and risk" published in issue 40 (3). Martin has been long associated with Business Information Review, and is one of the leading figures in commercial information management globally. The winning paper addresses the way in which employees use workarounds and shadow IT to cope with the complexity of enterprise applications, and is a deserving winner.

References

Berners-Lee, T. (1989), Information Management: A Proposal, available at: http://www.w3.org/History/1989/proposal.html, [Accessed 16th October 2021].

Berners-Lee, T. (1999), Weaving the Web: the Past, Present and Future of the World Wide Web by its Creator, London: Orion Business Press.

Brownlow J, Zaki M, Neely A, et al. (2015) *Data and Analytics - Data-Driven Business Models: A Blueprint for Innovation the Competitive Advantage of the New Big Data World*. Cambridge: Cambridge Service Alliance. Available at https://cambridgeservicealliance.eng.cam.ac.uk/system/files/documents/2015MarchPaperTheDDBMInnovationBlueprint.pdf (accessed: 12 February 2023)

Hess, A. (2008) Reconsidering the Rhizome: A Textual Analysis of Web Search Engines as Gatekeepers of the Internet

Muscolino, H., Machado, A., Rydning, J. & Vesset, D. (2023), *Untapped value: what every executive needs to know about unstructured data*, Needham, MA: IDC Research Ltd.

Sadowski J (2019) When data is capital: datafication, accumulation, and extraction. *Big Data and Society* Jan-June: 1–12.

Tanwar, M., Duggal, R. and Khatri, S.K., (2015), Unravelling unstructured data: A wealth of information in big data. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)* (pp. 1-6). IEEE

Tredinnick, L. & Laybats, C. (2023), Dangerous Data: analytics and information behaviour in the commercial world, *Business Information Review,* 40 (1): 10-20.

Tredinnick, L. & Laybats, C. (2024), Decarbonising Information Work, *Business Information Review,* 41 (1): 6-9.

White, M. (2023), Workarounds and Shadow IT – Balancing Innovation and Risk, *Business Information Review,* 40 (3): 114 – 122.