Data Storytelling on Multi-modal Knowledge Graph via Data Comics: a case study in Yanyuwa Language

Zhiping Liang[1], Zijie Zeng[1], Gloria Fernandez Nieto[1], Yuheng Li[1], Yi-Shan Tsai[1], Guanliang Chen[1], Zachari Swiecki[1], Dragan Gašević[1], John Bradley[2] and Lele Sha[1,*]

[1]*Center for Learning Analytics at Monash (CoLAM), Monash university*
[2]*Monash Indigenous Studies Centre, Monash university*

Abstract

This paper pioneers a study of storytelling on a multi-modal and multi-grained knowledge graph of the Yanyuwa language - a critically endangered Indigenous Australian language. Unlike traditional data management technologies (e.g., relational database), the knowledge graph is capable of connecting individual unit of entities (e.g., people, location or an abstract concept) to their cultural significance (e.g., oral traditions, historical context, examples of the concept), which is crucial for capturing the cultural and historical significance embedded in languages to support various stakeholders (e.g., indigenous educators) to revive the language. To ensure that the knowledge encoded can be accessible and understandable to non-technical stakeholders, who typically have no prior knowledge about the knowledge graph, this paper investigate a data comic approach to transform complex graphical knowledge into narratives that are easier to understand and more engaging for the non-technical audience. We report on: (i) technical details involved in constructing the knowledge graph; (ii) processes involved in transforming a graph segment into a narrated knowledge stories; and (iii) the design and prototyping process of the eventual data comic visualisations carried out by the three knowledge graph engineers. In doing so, we contributes novel insights regarding the role of data storytelling under the circumstances of employing a technically advanced solution like knowledge graph in the longstanding challenge of reviving endangered indigenous languages.

Keywords

Data storytelling, Multi-modal knowledge graph, Data comics

1. Introduction

> *"Nakari ridinja nungka rikarrangu nya-mangaji ni-maliji, kilha-nyngkarri nyinku wuka ngala yinda wukanyinjawu, yinbayawu ji-awarawu, barra bawuji wakara*

Joint Proceedings of LAK 2024 Workshops, co-located with the 14th International Conference on Learning Analytics and Knowledge (LAK 2024), Kyoto, Japan, March 18-22, 2024.
*Corresponding author.
zhiping.liang@monash.edu (Z. Liang); Zijie.Zeng@monash.edu (Z. Zeng); GloriaMilena.FernandezNieto@monash.edu (G. F. Nieto); yuheng.li@monash.edu (Y. Li); Yi-Shan.Tsai@monash.edu (Y. Tsai); Guanliang.Chen@monash.edu (G. Chen); Zach.Swiecki@monash.edu (Z. Swiecki); Dragan.Gasevic@monash.edu (D. Gašević); John.Bradley@monash.edu (J. Bradley); Lele.Sha@monash.edu (L. Sha)
0009-0003-9368-3460 (Z. Liang); 0000-0001-8967-5327 (Y. Tsai); 0000-0002-8236-3133 (G. Chen); 0000-0002-7414-5507 (Z. Swiecki); 0000-0001-9265-1908 (D. Gašević); 0000-0002-0356-6945 (J. Bradley); 0000-0002-8138-3853 (L. Sha)

ki-yabarri ni-wurdu, ngabaya jibiya baji ki-yibarra ni-maliji ajinjala yinku rdiyangu kurda kurdardi ji-wankalawu."

"From yesterday perhaps, or maybe early this morning those handprints, the spirit beings heard your words, when you spoke and sang for this country, so that was what they did, they felt good, so the spirit beings from this place put their handprints in the cave for you, they are newly made, they are not from a long time ago."

This opening quote comes from Mavis Timothy a-Muluwamara, a senior Yanyuwa woman, the Aboriginal owner of sections of land and waters throughout the southwest Gulf of Carpentaria, northern Australia. Like many indigenous languages, Yanyuwa demonstrate an incredibly intimate knowledge of their country. At the time of writing, Yanyuwa is one of the most critically endangered indigenous languages, with only two fluent speakers in the elderly group. Unlike the lost land, when you lose a language, a whole way of being, a whole cultural universe, is lost forever [1, 2]. The prior revitalisation effort has shown that the establishment of a knowledge repository (i.e., a data pool or language archive which contains critical indigenous knowledge including encyclopedia, books, etc) is a crucial step toward language revitalisation programme success [3, 4, 5, 6]. Since a sizeable knowledge repository effectively connects language with critical knowledge, enriching linguistic knowledge (e.g., prefix to words, grammar rules) with oral traditions (e.g., song, stories) and other spiritual / healing practice. Such cultural contextualisation of a language can contribute to strengthening cultural identity and enhancing the overall quality of language education [7, 5, 6].

While the establishment of these knowledge repositories represents a meaningful effort to preserve the critical knowledge about these languages, these repositories have rarely been tailored for practical utilisation, presenting a significant barrier for usage by the non-technical community members to adopt them for language revitalisation (e.g., designing language educational material) [8, 9, 10]. To tackle this, we aim to investigate the **R**esearch **Q**uestion:

> How can *indigenous knowledge* be represented and shared to foster *knowledge sharing* between technical (e.g., computational researchers) and non-technical stakeholders (e.g., indigenous educators, community leaders, and traditional knowledge and land owners)?

Inspired by data storytelling, this paper employed an innovative knowledge graph (KG) technology to construct such language repository. Different from prior approach [11, 12], the knowledge graph models the relations of different pieces of information in a graphical format, thereby providing semantically rich and deeply contextual knowledge about the Yanyuwa language. We argue that such "contextualisation" paves way for narrating more meaningful and engaging knowledge stories, which may be further polished with a "comic" layer to display visual knowledge stories to support non-technical stakeholders' sense making of knowledge encoded in the KG. In doing so, we seek to bring new insights to the evolving discourse on how to cultivate effective and inclusive language revitalisation practices for endangered indigenous languages [1].

[1]This project is approved by Monash ethics committee, Project ID: 39279

2. Related Work

2.1. Indigenous language revitalisation

Scholars have long recognized the urgency of preserving and reviving endangered indigenous languages, which are currently disappearing at a accelerated rate [13, 14]. Central to the revitalisation efforts has been the exploration of innovative methods that go beyond traditional linguistic documentation [8, 9, 15, 10]. Recent literature has increasingly emphasized the growing potential of advanced computational models, particularly in the natural language processing (NLP) domain [16]. For instance, in [17], the authors have systematically explored the potential use of state-of-the-art NLP tools to assist language education to revitalise the Cherokee language, including the use of automated question generation and language assessment material creation. In a similar vein, Teodorescu et al. [18] explore the use of various language models in the Cree context (an indigenous language spoken in North America), the authors concluded that NLP tools (e.g., text classifier) may be potentially adapted for community preferences and help with language revitalisation efforts. Despite these advances, researchers caution about the challenges of translating nuanced cultural contexts of indigenous languages into such digital formats [19, 20]. These insights from recent literature illustrate a growing recognition of the need for innovative, culturally sensitive methods in the documentation and revitalisation of endangered languages like Yanyuwa.

2.2. Knowledge graph

Recent development in computational techniques, such as knowledge graphs, open up new possibilities in reviving the indigenous language [21, 22]. Different from prior language revitalisation efforts which typically focuses on the documentation and archiving of linguistic knowledge [23], the knowledge graph is capable of modelling the interconnectedness between language and its cultural significance as a graphical representation. Conceptually, a knowledge graph G is composed of the ontology O, the data graph D and the corresponding relation R between them: G = (O, D, R), so the knowledge could be organised as the triples ("entity-relation-entity" or "entity-property-value") which could be scaled to a large graph network [24]. The adoption of knowledge graphs as a means to encode knowledge into practical knowledge bases has shown considerable promise, especially in the field of linguistics. In fact, the early establishment of knowledge graphs like WordNet [25] and ConceptNet [26] demonstrated a way to model and analyse English language by connecting words and concepts into an intricate web of semantic relationships, these graphs had subsequently been widely used to perform natural language understanding tasks [27, 28, 29]. In the context of language learning and education, the knowledge graph may also serve not just as repositories of linguistic knowledge (e.g., similar to a dictionary) but also as an intelligent tool to encapsulate and disseminate the cultural and contextual elements that are central to languages [30, 31, 32]. Although KG has potential to innovate the field of language revitalisation, such technical solutions are typically constructed by data scientists or technical personnel, and may be difficult to understand and utilise by non-technical stakeholders, such as indigenous educators, school policy makers or parents, whose roles are critically important to the success of a language revitalisation programme [33].

2.3. Data comics for data storytelling

To include and engage non-technical stakeholders in such KG-based language revitalisation initiative, we resorted to the growing body of research in human-computer interaction and data storytelling. Data storytelling is the combination of data, visuals and narrative to support casual users or users with less experience in complex data analysis scenarios [34, 35]. Data comics have emerged as an important visual storytelling branch to data storytelling by using sequential images, constructed from data-driven visualisations [36]. Fundamentally, a data comic has two dimensions: content relation and layout. Content relation describes a set of transitions between comic panels to construct a narrative. Comic layout is the organisation of comic panels. There are six types of content relations (i.e., narrative, temporal, faceting, visual encoding, granular, and spatial) and nine content layouts (i.e., large panel, annotated, grouped, tiled, parallel, grid, network, branching, and linear). These narrative components of data comics allow for a nuanced presentation of complex information, making them particularly useful in scenarios where understanding and participation of non-technical stakeholders are key [37]. Additionally, their visual nature supports the integration of diverse data types, from quantitative graphs (e.g., data points) to qualitative anecdotes (e.g., narratives), offering a comprehensive approach to data representation. In the work of [38] data comics was successfully used to explain diabetes in a complex and multi-modal healthcare context. In a similar vein, the work presented by [39] illustrated the potential of comics to explain complex networking data with temporal changes. Our research builds on the growing interest in data comics and their capabilities for effective multi-modal data communication in a language revitalisation context.

3. Methodology

3.1. Dataset

This study is based on the comprehensive Yanyuwa archive collected by John Bradley (one of the co-authors) over a 43 year period, which includes two volume dictionary, encyclopedia of the Yanyuwa language and digitised files of stories and songs composed and sung by contemporary Yanyuwa men and women, image/photograph/map collections about the lives of Yanyuwa people, narrative animations, diagrams and illustrations of material cultures. Table 1 provides a detailed overview of the source data in this collection.

3.2. Knowledge graph construction

To fully harness the capabilities of knowledge graphs for storytelling, we strategically encode knowledge at various levels of granularity and across multiple modalities. We argue that such diverse array of knowledge representation may open up possibilities for more compelling blend of visual elements that subsequently leads to better narration for engaging data comic stories.

Extracting Document and section entities. A team of three data science researchers (with no Yanyuwa expertise) worked collaboratively to systematically extracted knowledge entities of three granularity: document (i.e., a collection of text, audio file, or videos about the same topic, such as storybook, songbook, or a dictionary), sections (a meaningful segment of a document,

Table 1
An overview of the comprehensive archive adopted in this study, including songs, stories, encyclopedias, animations, documentaries, and recordings.

Type	Title	Description	Quantity	Authors/Contributors	Year
Text (contains image)	Songs from the gulf country	A collection of songs created by members of the Yanyuwa community, encompassing a range of themes such as traditions, the land, customs, rules, hunting practices, cuisine, geography, migration, nostalgia, and land disputes.	24 pages	Fiona Holley, Amanda Kearney, Yanyuwa community, and others.	2022
	Duwarra Wujara Kulu Bujimala: Two Young Men and the Rainbow Serpent	A story about the Journey of two young men through sacred sites, encountering the Rainbow Serpent, and learning cultural norms and consequences.	68 pages	Yanyuwa community, John Bradley, Amanda Kearney, Brent McKee, Craig Martin, and Others	2022
	A-barral kulu Nyamirningiya/A-barral kulu Mirningiya	Children's tale: "White Cockatoo and Man" in Yanyuwa language.	53 pages	Bella Charlie and Jemima Miller	1982
	Kiya-arri nya-ardu ngamala/ Ka-arri ardu ngamala	Children's tale: "A Boy in the South" in Yanyuwa language.	31 pages	Bella Charlie, Jemima Miller and Queenie Simon	1982
	A Yanyuwa Encyclopaedia Volume 2: Yanyuwa–English Dictionary	Part 2 of Volume 2 of the Yanyuwa Encyclopedia, serving as a Yanyuwa-English Dictionary, encompasses over 4,000 entries.	488 pages	Yanyuwa Families; John Bradley; Jean Kirton; Elfreda MacDonald	2017
	A Yanyuwa Encyclopaedia Volume 2: English–Yanyuwa Word Finder	Part 3 of Volume 2 of the Yanyuwa Encyclopedia, serving as an English-Yanyuwa word finder, encompasses over 4,000 entries.	129 pages	Yanyuwa Families; John Bradley; Jean Kirton; Elfreda MacDonald	2017
	Yanyuwa Wuka: Language from yanyuwa country	An Yanyuwa dictionary and culture resource, which could be considered as an earlier edition of the Yanyuwa encyclopedia.	636 pages	John Bradley, Jean Kirton and the Yanyuwa Community	1992
	A Yanyuwa Encyclopaedia Volume 1	Volume 1 of the Yanyuwa Encyclopedia, detailing the Yanyuwa community's culture, grammar, and traditional texts, including songs, poetry, and ancestral dreaming narratives.	593 pages	Yanyuwa Families; John Bradley; Jean Kirton; Elfreda MacDonald	2016
Text (contains no image)	A Yanyuwa Encyclopaedia Volume 2: Yanyuwa Grammar	Part 1 of Volume 2 of the Yanyuwa Encyclopedia, focusing on Yanyuwa grammar.	86 pages	Yanyuwa Families; John Bradley; Jean Kirton; Elfreda MacDonald	2017
	Further Aspects of the Grammar of Yanyuwa, Northern Australia	This book delves into more advanced aspects of Yanyuwa grammar.	230 pages	Jean Kirton and Bella Charlie	1996
Video	Duwarra Wujara Kulu Bujimala: Two Young Men and the Rainbow Serpent	The Journey of two young men through sacred sites, encountering the Rainbow Serpent, and learning cultural norms and consequences.	27 mins	Dinah Norman a-Manrrngaw, John Bradley, Brent D. McKee, and others	2022
	Turtle hunting, Tent boxing and Aeroplane Dance Documentary	This video records the activities of Yanyuwa people regarding their turtle hunting, tent boxing and Aeroplane Dance.	97 mins	Trevor Graham, Sharon Connolly, Cristina Pozzan, and others	1993
	Yanyuwa Culture Documentary and Journey East	A documentary about Yanyuwa people and their sea-based culture, focusing on dugong and turtle hunting, maintaining traditions despite modernization.	100 mins	Debbie Sonenberg, Yanyuwa comunity, and others	1989
	Songs from the gulf country	This animation includes various songs composed by people from the Yanyuwa community.	9 mins	Yanyuwa comunity, John Bradley, Sam Pankhurst, and others	2022
Audio	Audio Recordings	A collection of 134 recordings that showcase a rich tapestry of Yanyuwa culture, including interviews, storytelling sessions, dreaming narratives, traditional song performances, and conversations from the Yanyuwa people.	134 files	Yanyuwa community, John Bradley, Jean Kirton and others	—

such as a story within a storybook), and atomic knowledge (this may be a sentence, a phrase, or a video frame which cannot be meaningfully reduced further). Conceptually, these knowledge entities are "nodes" within a knowledge graph. Given the structured nature of our Yanyuwa dataset (e.g., a storybook containing chapters of stories), we can easily extract document and section entities via a manual segmentation (e.g., a single story within a storybook is extracted as a section knowledge entity). For audio and video data, text transcripts were used, and for images, captions were used. In section and atomic granularity, we excluded audio/video

without corresponding transcripts (around 62% of all audio/video data), since we cannot reliably extract meaningful textual representation (our exploration of using Whisper tool, which is a state-of-the-art speech recognition tool only reaches about 30% accuracy), we plan to include these data in the next phase of this study where we conduct in-depth interview sessions with Yanyuwa community knowledge owners and educators.

Extracting atomic entities and relations. For atomic entities, we resort to automated natural language processing techniques as each atomic entities may be of variable length, and cannot be easily segmented and extracted manually. The process is threefold. First, we conducted frequency analysis using Term Frequency-Inverse Document Frequency (TF-IDF) [40], which is a widely adopted text analytics approach to identify words / terms that are of high-importance in a document. Second, for all entities, we conducted dependency parsing of their "title" property (e.g., the name of a song) in the Yanyuwa dataset; this will extract the appearance of these entities title along with their most likely dependencies (e.g., a verb which connects two entities); this enables us to extract a large number of entity-relations in the graph. Third, to further establish missing relations, we conducted a co-occurrence analysis involving identifying the frequency with which entities appear together in a given dataset (not necessarily with a dependency relation). We measured the strength of association and, for top-100 strong association, we established the relation via manually analysing the surrounding text. A key motivation behind these entity-relation extraction approaches is so that all the encoded entities and relations are clearly traceable back to the source data. Alternatively, while other inference-based approaches (e.g., GPT) may be more powerful, we opt for not using them given that they may produce novel relations / entity names that are not traceable to the original Yanyuwa dataset and hence compromise the authenticity of the knowledge graph. Once all the entities and relations were established, we used Neo4j tool [2] to import them into a graphical format. A visual summative view of the KG can be seen in Figure 1.

3.3. Storytelling workshops on knowledge graph

We invited the three researchers who developed the KG in Section 3.2 to participate in the following design workshops. All three researchers are male, and have background in data science. Two of the three researchers have prior knowledge about KG ontology and constructions before participating this project. The team conducted six design workshop sessions in total.

Session 1: introduction to data comics. Given that KG researchers do not have prior expertise in data comics, we conducted an introductory session to introduce key concepts. In the first half-session, data storytelling fundamentals were introduced. Then, for the second half session, data comics were introduced, including two important comic dimensions: content relation and layout [41]. Details regarding six types of content relations and nine content layouts were introduced (see Section 2.3). An example data comic for each layout was presented to the participants.

Session 2: design objective. To help clarify the aim of the comic, prior to the comic design sessions, we had an in-depth discussion about how to communicate the knowledge encoded in KG with the non-technical stakeholders (e.g., indigenous educators), and how can data

[2]Neo4j (2023). Neo4j Aura. Neo4j, Inc. https://neo4j.com/cloud/platform/aura-graph-database/.

comic help address it. All three participants agreed that this is a complex question that may be more meaningfully tackled by first deconstruct it into smaller sub-question: i) How many nodes should be involved in a single data comic? ii) What are the key visual techniques / data storytelling principles in data comics that can help "narrate" a selection of nodes? iii) how should we mitigate the risk of compromising the authenticity of the original knowledge, given that each person may narrate the story differently and none of the participants were of Yanyuwa background? Each participant was given the opportunity to contribute to these questions via filling a sticky note, followed by a group discussion to resolve conflicts. The discussion resulted in following agreement and conflicts. We reached conclusion about sub-question i and ii. For sub-question i) although presenting KG nodes as a visual data story is context-specific, there should be a minimum number of node involved for a sufficiently informative story, all the node should be connected with at least one other node; sub-question ii) in data comics, the key visual knowledge is with regard to content relation, which can be mapped from KG entity-relation, after which, the comic layout could be determined. For sub-question 3, conflicting view were raised, on the one hand, one participant believe that the knowledge story narration should be based on KG-relations, therefore the visual stories should have a "ground truth" of a set of optimal narration. On the other hand, other participants believe that, despite being guided by the KG relation, comic design are still largely subjective given that each comic design decision (e.g., which knowledge to focus, and which connecting node to feature, which layout to adopt) will affect the final comic narration. As a compromise, the discussion moderator (an experienced data storytelling researcher) decided to alternatively discuss around a set of guidelines for designing a narrative structure based on perception theory [42, 43, 44, 45], which are recorded in a form of design **G**uidelines:

- G1 [Visual perception]: Visually explain and engage the audience about the multi-modal knowledge stories in the KG.
- G2 [Gestalt Principles]: Support recognition and sense-making via coherent and intuitive designs principles based on proximity and continuity of knowledge.
- G3 [Affordances]: Informing KG's use cases in language revitalisation and potentially guide the decision making about material design process.
- G4 [Cultural context]: Produce an accurate representation of the knowledge encoded by KG with traceability.
- G5 [Cultural context]: Explore the data comic's potential for designing engaging material for cultural revitalisation purposes.

Session 3: Sketching. To practice data comic design skills, a data comic researcher selected an example KG segment consisting of 8 KG entities about an animated Yanyuwa story "Mud crab", consisting of 3 atomic vocabulary entities, 3 atomic image entities, and 2 story text sections. The reason for selecting this example segment was due to that: it included a representative set of nodes from different type of entities and modalities. KG researchers were instructed to first think about the content relations of the entities for 20 minutes, then they selected comic

layout to sketch a comic that can best illustrate the KG segment based on design guidelines G1-5. Feedback was given by a senior data comic researcher, and a 30-min revision session were carried out to alter / polish the comic regarding: (i) if too little / too much information are presented in a single layout; (ii) if the comic simply does not make sense; (iii) if misleading visual cues were used which may cause confusion. We ensure that participants can design data comics independently at the end of this session.

Session 4: Selecting and creating KG stories. The three KG researchers were instructed to independently select a KG segment based on their judgement of content importance. The segment should contain at least 3 entities and should be interconnected.

Session 5: storyboard prototyping. For the selected KG segment, participants were instructed to independently design and construct comics using miro tool [3]. After all the participants finished prototyping. Participants were allowed to polish / or create an alternative comic about the same KG segment; however, only one comic was submitted per person for presentation during session 6.

Session 6: Presentation and discussion. To reflect on the design, we organised a discussion meeting for each participant to present their comic design and explain their design decisions. During presentations, in addition to explaining design decisions according to guidelines G1–5, more in-depth questions were asked to further our understanding about the KG comic design process: (i) rationale behind the selection of KG segment with respect to narration and whether entities should be included / removed to improve the narration; (ii) whether it is possible to automatically translate KG relation into content relation and comic layout (e.g., 'a Yanyuwa vocabulary from a song' may be translated to 'space-annotations' content relation, and be visually represented using Annotated layout with vocabulary in the peripheral and song in the middle panel); and (iii) the benefit / limitations of the use of multiple modalities (e.g., are certain modalities preferred over the other).

4. Result

4.1. Knowledge graph

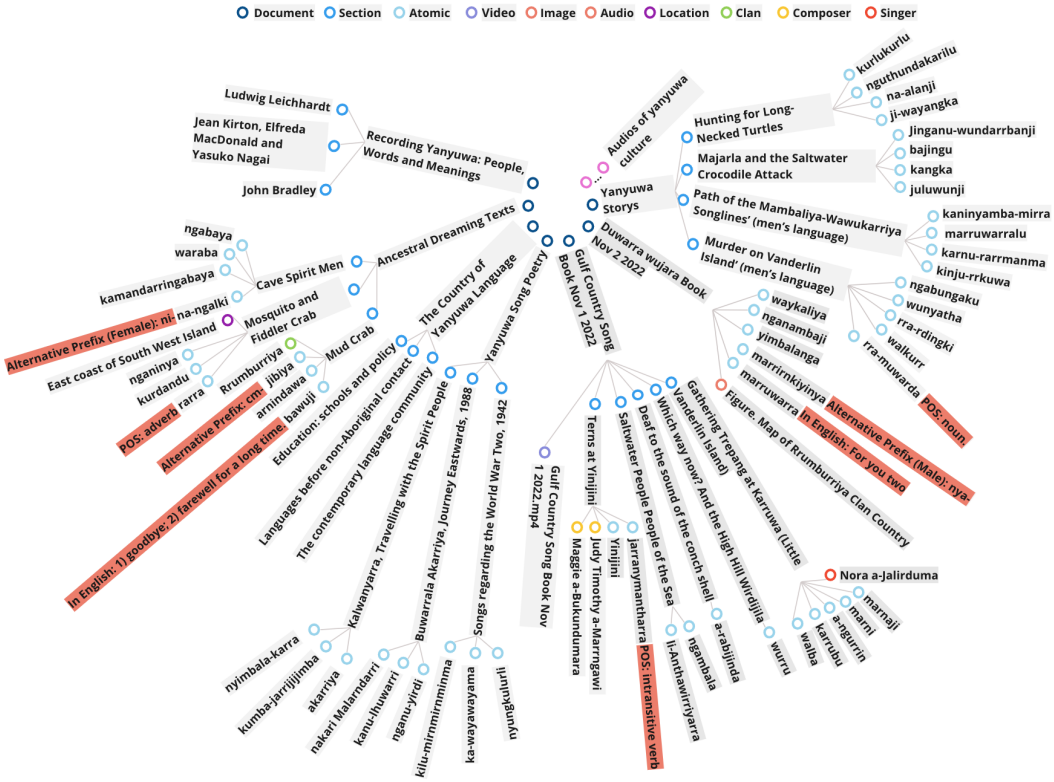
We present a summative view of KG containing all the representative entities of different types in Figure 1. The descriptive details of all entities (i.e., "Node (Granularity)" column), their properties ("Properties of Node"), possible relations with adjacent nodes ("Possible relations of a Node") and semantic category ("Possible Cattegory") extracted using TF-IDF (see Section 3.2), note that since atomic node (i.e., Granularity = 2) is a single atomic unit, they are automatically assigned a category based on their connected section node.

The central of Figure 1 are all the document type node (Granularity = 0), their immediately connected nodes are section type node (Granularity = 1), the node at the outer circle are atomic nodes (Granularity = 2). Logically, a document node can have many section node, and a section node can have many atomic node. Note that we have not exhaustively listed all the nodes and all the connections as it would be too much information to display / understand clearly, and

[3]Miro (2023). Miro online whiteboard. RealTimeBoard, Inc. www.miro.com.

instead, we opt for displaying the key representative node and their main structural information (i.e., the document-section-atomic structure).

Figure 1: An overview of our knowledge graph structure, different coloured nodes represent the unique type of each instance, the text in the grey box represents the instance name, and the text in the red box represents the attributes adjacent to the word node.



Upon analysis, we observed following advantages of this graphical way of organising knowledge: (i) **Traceability**. The hierarchical structure of document-section-atomic enables encoded knowledge to be efficiently traceable to the original source, e.g., when we aim to know about the events that occurred at a particular location, we begin by identifying the location node and then trace back through the 'Located at' or 'Happened at' relations to explore all the events linked to that place. Subsequently, we can further trace back through the event relations to uncover additional knowledge about these events and the source documents, thereby facilitating decision-making about the use case of the knowledge, for instance, the community may prefer including stories originated / authored by people of Yanyuwa background. Besides, any extensive use of the KG knowledge (e.g., in a language learning material) can be easily referred back to the KG entity via retrieval from KG (i.e., querying the knowledge name), thereby ensuring the authenticity of knowledge and enhancing accountability. (ii) **Alignment between different modalities**. We found that, by organising knowledge of different modality in an inter-connected manner, the knowledge of different modalities (i.e., audio, video, images, text) may be aligned semantically. For example, when identifying key atomic entities from video animations of

Table 2
Our knowledge graph consists of a total of six types of nodes, where the granularity level of each type of node belongs to one of the following: level **0** (the highest level), level **1** (the middle granularity level), or level **2** (the finest granularity level).

Node (**Granularity**)	Description	Possible Relations of Node	Properties of Node	Possible **Category**
Document (**0**)	The node of the highest granularity level (level **0**). Typically it could be a collection of sections (granularity level **1**) of a particular category (e.g., a collection of sections about multiple Yanyuwa stories).	Include; Authored by; Contributed by; Organized by; Edited by	Title; Year; Authors and Contributors; **Category**; Description; Topics and Keywords; Source File; Page Index in Source File; Source Stored at; Note	Song Poetry; General Yanyuwa Text; Ancestral Dreaming Text; Story; Reference (e.g., Grammar); Dictionary
Section (**1**)	The node of the granularity level **1** (finer than level **0**). Typically it could be several passages of text centering on a specific topic (e.g., a section could be centered around a particular Yanyuwa poetry).	Belong to; Include; Has; Composed by; Written by; Sung by; Related to	Title; Year; Authors and Contributors; **Category**; Description; Source File; Page index in Source File; Source Stored at; Note	Song Poetry Text; Article Text; Background and Information Text; Ancestral Dreaming Text; Story Text; Other Text
Location (**2**)	The node of the finest granularity level **2**, particularly used for the entity of location.	Belong to; Located at; Related to; Happened at; Part of	Title; Description	
Event (**2**)	The node of the finest granularity level **2**, particularly used for the entity of event.	Happened at; Related to; Presented by; Collaboration between	Title; When; Where; Description	
Vocabulary (**2**)	The node of the finest granularity level **2**, used for the entity of the general Yanyuwa words.	Synonym; Has; type of; part of; associated with	Title; Explanation; Parts of Speech	
Person (**2**)	The node of the finest granularity level **2**, used for the entity of a particular person.	Wrote; Composed; Sang; Is / Was; Said	Title; Description	

background stories of 'Gulf Country Song', similar atomic entities can often be extracted from the original song text appears in the book, thereby established a one-to-one connection between song lyric and frames of video. In more complicated cases, an image about Hammerhead shark appear in encyclopedia book can also be associated with a story in the storybook about the shark with key vocabulary terms explained in the dictionary. (iii) **Knowledge association and connection**. By qualitatively analysing nodes, we are able to infer latent relations between nodes which further enhance the semantics of the encoded knowledge. For example, when we examine the cooking Turtle series, we can deduce that some of the hunting or cooking tools might be potentially associated with this particular event and then we can try to explore these potential connections to broadening the scope of knowledge encompassed by it. This is further exemplified and detailed in our data comic result, where the multi-modal comic elements may appear from multiple sources, see Figure 4.

While KG constructed in this way has clearly demonstrated advantages in terms of content narration, we note that the current version of KG are constructed by data scientist and have not yet involved Yanyuwa community member to validate its correctness and comprehensiveness, which is planned to be the next step of our study. Additionally, we also find that as KG scales (i.e., by adding more entity-relation-entity triple), the validation of KG quality (e.g., certain connection such as "related" may not be specific enough) becomes more difficult. Indicating a need for exploring automated KG validation approaches.

4.2. Narrating knowledge stories

In preparation to prototyping data comics, we have conducted storytelling workshop session 1–3, from which the three KG researchers gained: (i) introductory knowledge about comics [session 1]; (ii) the goals of what and how comic should convey key KG insights to the audience based on a set of explicit guidelines, see Section 3.3 [session 2]; (iii) practical experience in comic design and drawing via sketching [session 3]. We hereby continue to reporting the results in session 4–5 about the selection and narration decision of a selected segment of KG, and how they mapped to comic design based on the pre-determined guidelines. See Table 3, the KG segments selected by the three participants are reported in column "Selected KG segments", the narration process is reported in column "Narration rationale" and the chosen data comic relation (a.k.a "content relation" as termed in the original data comic paper) is reported in "Comic relation".

Table 3
An overview of the decision justification in this study includes the relations that can be translated from content to comic, an overview of the selected KG segment, the narration process, the use of multiple modalities and the corresponding participant

Comic relation	Selected KG segments	Narration rationale	Modalities				Participant
			Text	Image	Video	Audio	
Temporal, narrative	Section node: Gulf Country Song Book Nov 1 2022, Gulf Country Song Book Nov 1 2022.mp4, Terns at Yinijini, Judy Timothy a-Marrngawi, Maggie a-Bukundumara, Map. Atomic node: Jarranymantharra, wangarrabantharra , Yinijini.	I have chosen this because I think it can help the narration process by utilizing the word nodes linked to the song contain essential linguistic knowledge necessary to explain the yanyuwa word. Additionally, the inclusion of an audio file enhances the conveyance of its linguistic knowledge. The song-related image files contain the composer's photo, and when combined with artistic singing animations in the video, these may facilitate the conveyance of social knowledge.	✔	✔	✔	✔	Participant 2 (P2)
Narrative	Section node: A Yanyuwa Encyclopaedia Volume 1, A Yanyuwa Encyclopaedia Volume 2, Ancestral Dreaming Texts, Hammerhead Shark, Map. Atomic node: Warriyangalayawu, adumu, ngurdungurdu	To explain the words in the article 'Hammerhead Shark', I retrieved associated words through a KG. And aiming to enhance the data comic's presentation, I incorporated illustrations from Yanyuwa Encyclopaedia Volume 2 based on these connections. Considering the travel routes mentioned in the stories included in the article, I included a map to help present the potential environmental knowledge.	✔	✔			Participant 3 (P3)
Temporal, narrative	Section node:A Yanyuwa Encyclopaedia Volume 2, turtle cooking series image 1, turtle cooking series image 2, turtle cooking series image 3, turtle cooking series image 4. Atomic node: ngunduwamantharra, na-marnda, wunakaka, wubantharra .	I selected four illustrations that vividly represent Yanyuwa culture to serve as the foundation for presenting social knowledge. Additionally, I included word nodes associated with these illustrations to explain the linguistic knowledge embedded in the Yanyuwa vocabulary.	✔	✔			Participant 1 (P1)

Overall, we observed that participants generally selected different contexts for narration including: turtle cooking [P1], song [P2], and shark encyclopedia [P3] respectively. The contexts selection was mainly based on: (i) the availability of imagery [P1, P2, P3] and video / audio [P2] modalities for engagement purposes (G1); and (ii) perceived environmental and social significance (G4). Though not explicitly, all participants followed the rule of proximity and continuity of knowledge (G2) and opted for knowledge within a single context. We particularly draw attention to P3, where the map node were extracted from Encyclopedia to complement the Hammerhead shark story in Ancestral dreaming text, which enriched the story with a dynamic location context (i.e., the travels over tag in the comic Figure 4). All participants emphasized the importance of linguistic knowledge in their respective narrative (G3), and subsequently selected atomic / section nodes from Yanyuwa dictionary to highlight the: (i) part of speech (e.g., noun, verb) [P1, P2, P3]; (ii) English definition [P1, P2, P3]; and (iii) Synonym [P2];

Figure 2: Data Comic for the Turtle Cooking series with its corresponding KG segment, which is drwan by P1.

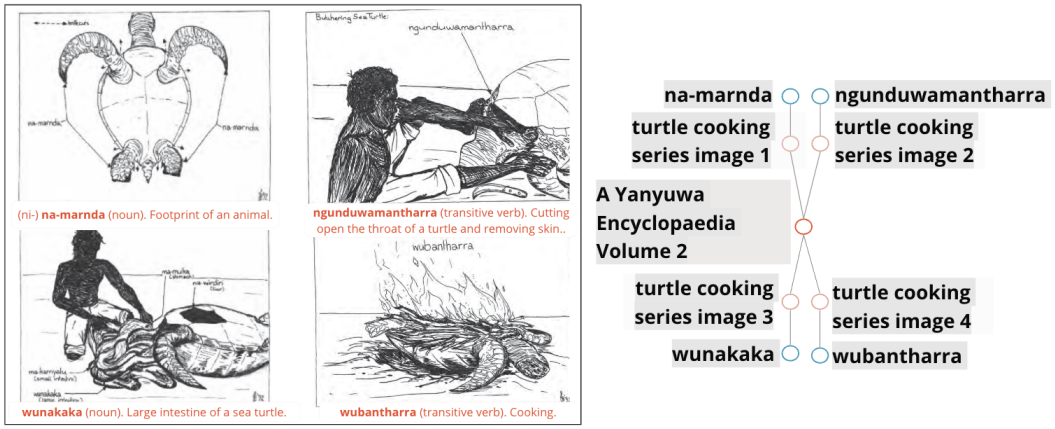
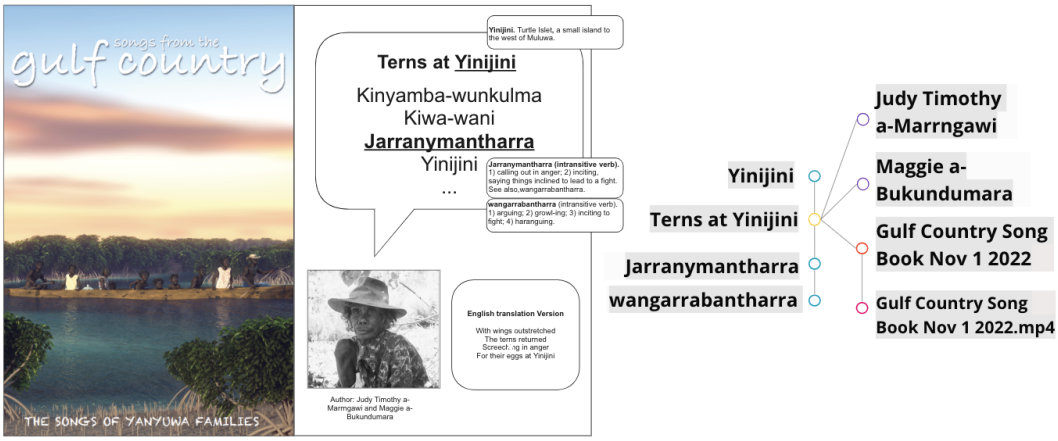


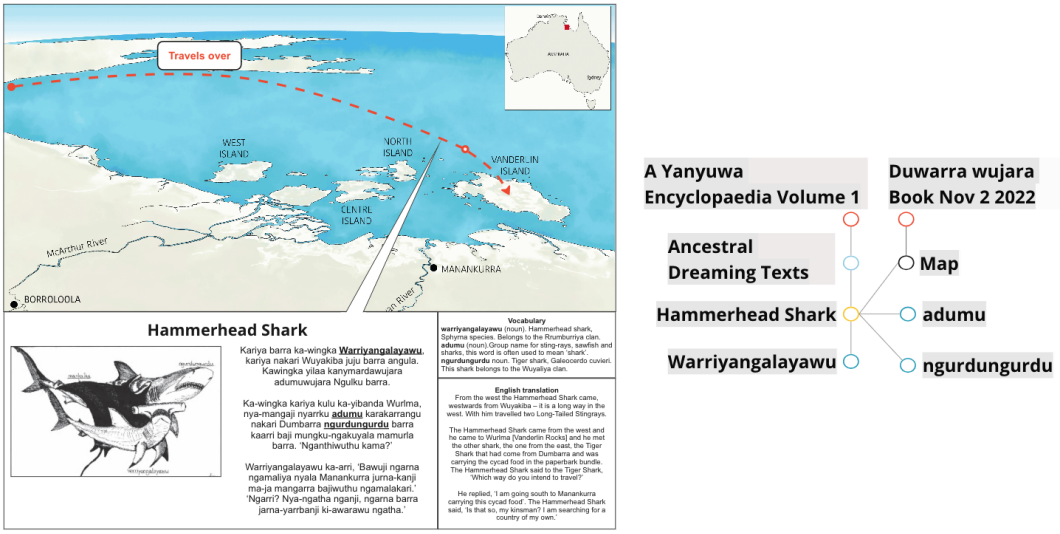
Figure 3: Data Comic for the Yanyuwa Song 'Terns at Yinijini' with its corresponding KG segment, which is drwan by P2.



4.3. Data comics

In this section, we report three final comic design results obtained at Session 6 for three selected KG segment in Table 3, respectively, as shown in Figure 2, Figure 3 and Figure 4. Note that the comic is displayed in the left, while a visualised KG selection is displayed to the right for traceability purposes (G4). On average, each comic displays 4–8 panels. We highlight key observations in following aspects: (i) **Narrative complexity**. We observed a clear difference in level of complexities in narrating on different selections of KG segment, which have been highlighted during the comic presentation time. Specifically, all participants agreed that including multiple modalities does not necessarily lead to narration complexity (as in Figure 1), and, P1 (Figure 2) is the least complex while P3 (Figure 4) is the most complex. We posit that this is related to the KG segment selection as guided by the proximity and continuity of knowledge

Figure 4: Data Comic for the Hammerhead Shark Story with its corresponding KG segment, which is drwan by P3.



(G2). P2, for example noted that "if removing song translations and vocabulary panels, the comic would be cleaner and easier to understand" when questioned about the inclusion / exclusion of the nodes. (ii) **Content relation and translation**. Generally, all participants agreed that it is possible to extract a set of generalisable rules for translating KG relation into content relation, for instance, KG nodes that are tagged with sequential KG relations (e.g., "followed by") could be mapped to temporal or spatial content relations i.e., Figure 3. Similarly, relations such as authored by, composed by may be mapped to narrative. However, given that 1 KG relation may be mapped to multiple content relation, and 1 content relation may be mapped to multiple comic relation, this indicated that more constraints needs to be placed to facilitate automated KG to comic translation. (iii) **Composite layouts**. Apart from P1, both P2 and P3 contains multiple layouts, i.e., P2 includes "grouped" (overall) and "tiled" layouts, while P3 includes "Large panel" and "tiled". However, a trade-off between clarity in overall visual stories (G1) and the sense-making of individual knowledge panels (G2) may need to be made. As P2 noted that "I wish to add important Yanyuwa English translations as annotations, and potentially the more detailed definition as a new branch using branching layout, but it becomes overly complicated." So, while adding translations and additional vocabulary may help with understanding the details of the song, the attention may be drifting away from the song to the branching panels, it is for this reason that the vocabularies nodes are ultimately added as a small tile so that the main attention is within the song content. A possible solution to mitigate this challenge is via interactive data comics [46], i.e., the element may dynamically appear only if user's mouse hovers over.

5. Discussion and Future Work

In this paper, we have pioneered a study combining the power of KG with storytelling, and situated this study in the important domain of indigenous language revitalisation. We have investigated the use of data comic to generate visual insight about a Yanyuwa knowledge graph. We conducted design sessions and reported our design process and final comic design.

5.1. Implications

The main implications of this study is threefold. First, given the unique advantages of knowledge association in the KG, this may enable rapid processing of narration which translate the low-level KG relation into more meaningful high-level content relation (as demonstrated in Table 3). For instance, we may extract all the KG segments that are relevant to key knowledge priorities as determined by Yanyuwa community, and use that as a basis to co-design language learning materials with Yanyuwa language educators. The traceability of KG means that we may adopt an iterative approach which may alter the narration past the initial stages (e.g., using a different node in the KG instead), and revise based on feedback obtained from different community members / knowledge owners. Second, we also find that all the KG relations are translatable into content relations (i.e., temporal and narrative). Although, such translation may not be one-to-one, e.g., the hammerhead shark series [P3] may also be identified as granular content relation between the map overview, detailed map location, and appearance of hammerhead shark. This implies that, we could potentially adopt a new approach to construct KG, by designing a subset of more fine-grained content relations to connect entities in KG. Therefore, KG is connected using content relation directly and may be more directly mapped into a comic layout to support rapid and automated prototyping of comics. As part of the future work, we may, for example, display automatically-generated comics that are highly ranked to be of high quality to other non-technical stakeholders, while revising the low-quality comics, rather than creating comics from scratch. Third, we noted a potential trade-off between understanding global (the main knowledge conveyed by the comic, such as the "Terns at Yinijini" song content) and local knowledge (the explanatory knowledge to facilitate the understanding of the main knowledge, such as the vocabulary explanation), while the more information added can enhance local information understanding, the global attention may be shifted to undesirable panels. This implies that, comics displayed may need to consider reader's prior knowledge, so as to avoid display excessive panels, e.g., displaying related knowledge panels for all the vocabularies which may lead to information overload.

5.2. Limitations and future work

We acknowledge following limitations. First, our KG, though being carefully constructed and validated within the internal team, has not undergone validation by the Yanyuwa community. The main reasoning is due to the critically endangered status of the language, and of the few people who can still speak and understand the language (including John Bradley who is the paper co-author, and two elderly Yanyuwa women), they lack the technical expertise in knowledge graph. We hope the adoption of data comics can facilitate this process, and help us communicate

the KG with the non-technical stakeholders to establish a meaningful collaboration. Second, due to copyright concerns, we have only included in our KG the data collected by John Bradley (co-author of the paper) over a 43 period of time, we acknowledge that there are other sources which may contain meaningful Yanyuwa knowledge data such as governmental digital repositories including The Australian Institute of Aboriginal and Torres Strait Islander Studies. Our KG is evolving and we will add new knowledge entities on a continual basis.

6. Acknowledgments

This research was supported by Google 2023 Award for Inclusion funding and Monash Data Future Institute (MDFI) 2023 seed funding grant.

References

[1] J. Fishman, What do you lose when you lose your language? (1996).

[2] J. J. Bradley, Can my country hear English?: Reflections on the relationship of language to country (2018). URL: https://bridges.monash.edu/articles/journal_contribution/Can_my_country_hear_English_Reflections_on_the_relationship_of_language_to_country/6025808. doi:10.4225/03/5ab86f0221aa1.

[3] T. S.-K. Marja-Liisa Olthuis, Suvi Kivelä, Revitalising indigenous languages: How to recreate a lost generation, volume 10, Multilingual Matters, 2013.

[4] P. Ciwas, Indigenous language education in taiwan (2009).

[5] D. Stiles, Four successful indigenous language programs., 1997. URL: https://api.semanticscholar.org/CorpusID:150552256.

[6] M. Walsh, 2 why language revitalisation sometimes works, 2010. URL: https://api.semanticscholar.org/CorpusID:154220552.

[7] B. Harrison, R. Papa, The development of an indigenous knowledge program in a new zealand maori-language immersion school, Anthropology and Education Quarterly 36 (2005) 57–72. URL: https://hdl.handle.net/10289/1359. doi:10.1525/aeq.2005.36.1.057, journal Article.

[8] P. Eisenlohr, Language revitalization and new technologies: Cultures of electronic mediation and the refiguring of communities, Annual Review of Anthropology 33 (2004) 21–45. URL: https://doi.org/10.1146/annurev.anthro.33.070203.143900. doi:10.1146/annurev.anthro.33.070203.143900. arXiv:https://doi.org/10.1146/annurev.anthro.33.070203.143900.

[9] I. Wagner, New technologies, same ideologies: Learning from language revitalization online, Language Documentation & Conservation 11 (2017) 133–156. URL: https://api.semanticscholar.org/CorpusID:157824290.

[10] C. Bowern, The Oxford Guide to Australian Languages, Oxford University Press, 2023. URL: https://doi.org/10.1093/oso/9780198824978.001.0001. doi:10.1093/oso/9780198824978.001.0001.

[11] G. Auld, The role of the computer in learning ndj bbana, Language Learning and Technology 6 (2002).

[12] C. Bowern, Chirila: Contemporary and historical resources for the indigenous languages of australia 10 (2016) 1–44.
[13] N. H. J. W. D. C. N. H. Lenore A. Grenoble, Dartmouth College, Saving Languages: An Introduction to Language Revitalization, 2005.
[14] S. Romaine, Preserving endangered languages, Language and Linguistics Compass 1 (2007) 115–132. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2007.00004.x. doi:https://doi.org/10.1111/j.1749-818X.2007.00004.x. arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-818X.2007.00004.x.
[15] C. K. Galla, Indigenous language revitalization, promotion, and education: function of digital technology, Computer Assisted Language Learning 29 (2016) 1137–1151. URL: https://doi.org/10.1080/09588221.2016.1166137. doi:10.1080/09588221.2016.1166137. arXiv:https://doi.org/10.1080/09588221.2016.1166137.
[16] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimedia Tools and Applications 82 (2023) 3713–3744. URL: https://doi.org/10.1007/s11042-022-13428-4. doi:10.1007/s11042-022-13428-4.
[17] S. Zhang, B. Frey, M. Bansal, How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1529–1541. URL: https://aclanthology.org/2022.acl-long.108. doi:10.18653/v1/2022.acl-long.108.
[18] D. Teodorescu, J. Matalski, D. Lothian, D. Barbosa, C. Demmans Epp, Cree corpus: A collection of nêhiyawêwin resources, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6354–6364. URL: https://aclanthology.org/2022.acl-long.440. doi:10.18653/v1/2022.acl-long.440.
[19] K. Onyenankeya, Indigenous language newspapers and the digital media conundrum in africa, Information Development 38 (2022) 83–96. URL: https://doi.org/10.1177/0266666920983403. doi:10.1177/0266666920983403. arXiv:https://doi.org/10.1177/0266666920983403.
[20] M. Mager, X. Gutierrez-Vasques, G. Sierra, I. Meza-Ruiz, Challenges of language technologies for the indigenous languages of the Americas, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 55–69. URL: https://aclanthology.org/C18-1006.
[21] M. I. Huilcán Herrera, The use of technologies in language revitalisation projects: Exploring identities, Journal of Global Indigeneity 6 (2022).
[22] C. K. Galla, Indigenous language revitalization, promotion, and education: function of digital technology, Computer Assisted Language Learning 29 (2016) 1137–1151. URL: https://doi.org/10.1080/09588221.2016.1166137. doi:10.1080/09588221.2016.1166137. arXiv:https://doi.org/10.1080/09588221.2016.1166137.
[23] K. Thorpe, M. Galassi, Rediscovering indigenous languages: The role and impact of libraries

and archives in cultural revitalisation, Australian Academic & Research Libraries 45 (2014) 81–100. URL: https://doi.org/10.1080/00048623.2014.910858. doi:10.1080/00048623.2014.910858. arXiv:https://doi.org/10.1080/00048623.2014.910858.

[24] L. Tian, X. Zhou, Y.-P. Wu, W.-T. Zhou, J.-H. Zhang, T.-S. Zhang, Knowledge graph and knowledge reasoning: A systematic review, Journal of Electronic Science and Technology 20 (2022) 100159. URL: https://www.sciencedirect.com/science/article/pii/S1674862X2200012X. doi:https://doi.org/10.1016/j.jnlest.2022.100159.

[25] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, Language, Speech, and Communication, MIT Press, Cambridge, MA, 1998.

[26] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4444–4451.

[27] M. Barbouch, S. Verberne, T. Verhoef, Wn-bert: Integrating wordnet and bert for lexical semantics in natural language understanding, Computational Linguistics in the Netherlands Journal 11 (2021) 105–124. URL: https://www.clinjournal.org/clinj/article/view/130.

[28] R. Mao, C. Lin, F. Guerin, Word embedding and WordNet based metaphor identification and interpretation, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1222–1231. URL: https://aclanthology.org/P18-1113. doi:10.18653/v1/P18-1113.

[29] Z. Zhang, X. Geng, T. Qin, Y. Wu, D. Jiang, Knowledge-aware procedural text understanding with multi-stage training, in: Proceedings of the Web Conference 2021, WWW '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 3512–3523. URL: https://doi.org/10.1145/3442381.3450126. doi:10.1145/3442381.3450126.

[30] F. Sun, M. Yu, X. Zhang, T.-W. Chang, A vocabulary recommendation system based on knowledge graph for chinese language learning, in: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), 2020, pp. 210–212. doi:10.1109/ICALT49669.2020.00068.

[31] W. Chansanam, Y. Jaroenruen, N. Kaewboonma, K. Tuamsuk, Culture knowledge graph construction techniques, Education for Information 38 (2022) 233–264. URL: https://doi.org/10.3233/EFI-220028. doi:10.3233/EFI-220028, 3.

[32] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, Arco: The italian cultural heritage knowledge graph, in: C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon (Eds.), The Semantic Web – ISWC 2019, Springer International Publishing, Cham, 2019, pp. 36–52.

[33] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review 56 (2023) 13071–13102. URL: https://doi.org/10.1007/s10462-023-10465-9. doi:10.1007/s10462-023-10465-9.

[34] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H. J. Schulz, M. Streit, C. Tominski, Characterizing guidance in Visual Analytics, IEEE Trans. Visualization & Comput. Graph. 23 (2017) 111–120. doi:10.1109/TVCG.2016.2598468.

[35] C. N. Knaflic, Storytelling with data: A data visualization guide for business professionals, 12 ed., John Wiley Sons, New Jersey, 2017.

[36] Z. Zhao, R. Marr, N. Elmqvist, Data comics: Sequential art for data-driven storytelling,

HCIL Technical Report (2015).
[37] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, N. Henry Riche, Design patterns for data comics, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–12. URL: https://doi.org/10.1145/3173574.3173612. doi:10.1145/3173574.3173612.
[38] S. McNicol, Humanising illness: presenting health information in educational comics., Medical humanities (2014). doi:10.1136/medhum-2013-010469.
[39] B. Bach, N. Kerracher, K. W. Hall, S. Carpendale, J. Kennedy, N. Henry Riche, Telling stories about dynamic networks with graph comics, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 3670–3682. URL: https://doi.org/10.1145/2858036.2858387. doi:10.1145/2858036.2858387.
[40] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.
[41] B. Bach, Z. Wang, M. Farinella, D. Murray-Rust, N. Henry Riche, Design patterns for data comics, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–12. URL: https://doi.org/10.1145/3173574.3173612. doi:10.1145/3173574.3173612.
[42] J. J. Gibson, The Ecological Approach to Visual Perception, Houghton Mifflin, Boston, 1979.
[43] K. Koffka, Principles of Gestalt Psychology, Harcourt, Brace, New York, 1935.
[44] R. V. Almeida, R. Woods, T. Messineo, R. Font, The cultural context model: An overview., 1998. URL: https://api.semanticscholar.org/CorpusID:210466470.
[45] J. J. Gibson, The theory of affordances, in: J. B. Robert E Shaw (Ed.), Perceiving, acting, and knowing: toward an ecological psychology, Hillsdale, N.J. : Lawrence Erlbaum Associates, 1977, pp. pp.67–82. URL: https://hal.science/hal-00692033.
[46] Z. Wang, H. Romat, F. Chevalier, N. H. Riche, D. Murray-Rust, B. Bach, Interactive data comics, IEEE Transactions on Visualization and Computer Graphics 28 (2022) 944–954. doi:10.1109/TVCG.2021.3114849.