Making the Model Visible: Modelling Performance and Methodological Provenance in an AI-Assisted Digital Humanities Workflow

Sanjin Muftić
University of Cape Town
sanjin.muftic@uct.ac.za

Abstract

The South African College of Music (SACM) at the University of Cape Town holds a substantial collection of concert recordings and accompanying programme data. Preparing this material for publication through Ibali: Digital Collections UCT required relationships implicit in the original spreadsheets to be represented explicitly within Omeka S.

This paper examines the development of a linked-resource model structured around four resource types: Concert, Performance, Work and Performance Credit. Particular attention is given to distinguishing a musical Work from an individual Performance, and to representing performer participation as a separate resource when information such as instrument, voice type or character belongs to the relationship itself. The paper also considers practical consequences of the model, including the use of stable identifiers across successive imports and the challenge of creating meaningful labels for relational resources within the user interface.

The model was developed iteratively with generative AI used to explore structures, generate data-transformation methods and test possible representations. Drawing on this process, the paper argues that data modelling makes interpretive distinctions computationally visible, while significant human-AI modelling decisions should themselves form part of the methodological provenance through which digital humanities resources can later be understood and reused.

1 Introduction

The South African College of Music (SACM), founded in Cape Town in 1910 and incorporated into the University of Cape Town in 1923, has a long history of music teaching and performance.[1] Over this history, SACM has accumulated recordings of concerts, competitions and other performances, accompanied by programmes containing contextual information about the works performed, their composers, the musicians involved and, in many cases, their instruments or roles. Recordings and programmes have largely been retained on institutional file storage, although an earlier SACM digital project made a small selection of recordings available online.[2] That project organised material primarily around concerts, with individual performances represented through descriptive metadata. The current SACM Recording Repository project returns to this material at a larger scale and considers how programme and recording data can be structured through explicit relationships between Concerts, Performances, Works and Performance Credits. The SACM Recording Repository project emerged from a request to UCT Libraries to consider how this collection might be made more accessible and discoverable online. For the initial repository pilot, SACM provided programme data from 2025, with the modelling exercise limited to evening concerts.

SACM had already undertaken considerable work to capture information from the programmes in a spreadsheet. Individual performed pieces were represented as rows containing information about the work, composer and performers, while merged cells visually grouped performances belonging to the same concert. The spreadsheet was therefore largely understandable to a human reader, who could infer relationships from its layout. Those relationships, however, were not always explicitly represented in the data itself.

UCT Libraries' existing platform for this work

[1] For a fuller history of the South African College of Music, including its founding in 1910 and incorporation into the University of Cape Town in 1923, see the SACM institutional history page: https://humanities.uct.ac.za/college-music/history-0.

[2] See the earlier SACM digital collection: https://ibali.uct.ac.za/s/sacm/page/index.

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

was Ibali: Digital Collections UCT[3], a university-wide digital collections showcase built on Omeka S[4]. Omeka S was particularly suited to this experiment because individual items can be described through configurable resource templates and linked to other resources within the repository. This made it possible to represent explicit relationships between concerts, performances and works within the programme data. Before the SACM data could be imported into Omeka S, decisions therefore had to be made about the entities represented in the collection, the metadata used to describe them, and the relationships between them.

These decisions were also shaped by the question of findability. The source collection largely consists of recordings of complete concerts, but a user may instead be looking for a particular composition, composer, performer, instrument or genre. Treating the concert recording as the principal descriptive object would preserve the organisation of the existing files while providing limited access to the individual performances contained within them. A more granular model could describe a concert as containing multiple performances, distinguish each performance from the musical work being performed, and associate musicians with their participation in a particular performance. This also allows the descriptive structure to remain more granular than the available media, leaving open the possibility of attaching individual performance files in the future.

Questions of copyright, recording and dissemination permissions, access, storage, preservation and the manual labour required to capture and enrich programme metadata remain important to the wider repository project but fall outside the main scope of this paper. This paper focuses on the modelling of the descriptive data and the relationships through which the collection can be explored.

The data model was not determined fully in advance. It developed through the cleaning and examination of the source data, experimentation with possible structures, and a series of conversations with the generative AI system ChatGPT. AI was used to explore possible representations and to generate computational methods for restructuring the data. Proposed solutions were nevertheless tested against the collection and, in some cases, modified or rejected when they failed to represent distinctions apparent in the source material. Two decisions became particularly significant: distinguishing a musical Work from a particular Performance of that work, and creating a separate Performance Credit through which a performer and their role could be associated with an individual performance.

Using these decisions as its central case, this paper examines data modelling as an interpretive Digital Humanities practice and considers how generative AI participated in that process. It argues that data models shape which distinctions and relationships remain visible computationally, and therefore influence the forms of discovery and research a digital collection can support. It further asks whether significant moments in which AI-assisted representations are proposed, questioned and revised should form part of the methodological provenance through which the resulting model can later be understood, evaluated and reused.

2 From Programme Data to a Linked-Resource Model

Moving the SACM data into Omeka S required the relationships implicit in the original spreadsheet to be represented explicitly. A single programme row could contain information about a concert, the musical work being performed, the particular performance of that work, and the musicians involved. Treating the row itself as the primary object would therefore collapse several different kinds of information into a single record.

Here, linked-resource model describes the explicit relationships created between Omeka S resources within the repository. The term is used here for the internal structure of the collection and does not imply a broader Linked Data publishing framework. The model that emerged separated the data into four principal resource types: Concert, Performance, Work and Performance Credit. A Concert represents the larger event within which performances take place. A Performance represents the occurrence of a musical work within a particular concert. A Work represents the musical composition independently of any individual performance. Performance Credit records the participation of a performer or ensemble in a specific Performance and can retain information associated

[3]Ibali: Digital Collections UCT is UCT Libraries' digital collections platform for publishing and showcasing cultural, research and heritage collections: https://ibali.uct.ac.za/.

[4]Omeka S is an open-source web publishing platform designed for institutions managing and presenting linked digital collections: https://omeka.org/s/.

with that participation, such as instrument, voice type or character.

For the 2025 evening-concert pilot, this restructuring produced 24 Concert records, 330 consolidated Works, 364 Performances and 632 Performance Credits, resulting in 1,350 resources across the four resource types. Producing these resources involved separating repeated concert-level information, consolidating recurring Works, and restructuring performer, instrument and role information into individual Performance Credits.

The conceptual relationships were translated into four Omeka S resource templates, with the principal relationships represented as follows:

These resource types were implemented in Omeka S through separate resource templates for Concert, Performance, Work and Performance Credit. The model was developed within the existing vocabulary environment used by Ibali. Schema.org provided most of the classes and relational properties needed for the repository, supplemented where appropriate by Dublin Core and other vocabularies already available on the platform. Concert and Performance were represented as `schema:MusicEvent`, Work as `schema:MusicComposition`, and Performance Credit as `schema:Role`. This provided enough structure for the immediate requirements of the repository while keeping the model manageable within the existing Omeka S environment.

The trade-off is that some distinctions represented more formally in specialist music models are handled here through a smaller set of general-purpose properties and project-specific resource templates. Introducing a specialist music ontology would also have implications beyond this pilot, since vocabularies added to Omeka S become part of the wider metadata environment available across the platform. Adopting one would require additional work around implementation, documentation and ongoing maintenance. For an initial collection-specific modelling exercise, that additional commitment was difficult to justify.

Implementing these relationships also required a structure of stable local identifiers. Because the different resource types were imported into Omeka S in separate stages, later imports needed to refer reliably to resources created earlier. Works, Concerts and Performances were therefore assigned local identifiers that could function as linking keys between batches. Omeka S also assigns each imported resource its own internal numeric ID, but these system-generated identifiers provide little information to someone working with the source data outside the platform. The locally assigned identifiers therefore also followed recognisable patterns that made the resource type and, in some cases, its relationship to a Concert easier to identify during data preparation and troubleshooting. For example, `2025-04-08_CON06_P004` identifies a Performance associated with Concert `2025-04-08_CON06`. These identifiers were designed primarily to support the internal workflow and should not be understood as globally persistent identifiers or external URIs.

Table 1 summarises the principal classes and relationships used in the four resource templates. Having stable identifiers solved the linking problem, but it raised another question: what should these resources actually be called when they appear on the site? Concerts and Works generally had titles that could be carried directly into the repository, but Performance and Performance Credit were relational resources without an obvious standalone name. Their titles needed enough context to make sense when displayed elsewhere on the site, particularly within Omeka S's Linked Resources lists. The current Performance titles therefore combine information about the Work, Concert and date, while Performance Credit titles identify the participant and Performance. These can become lengthy, creating a tension between what works for the underlying data model and what makes sense to someone navigating the site.

This question becomes particularly visible through Omeka S's Linked Resources display. A Concert page may list all of the Performances that link back to it, while a Performance may similarly expose its associated Performance Credits. The title therefore does more than identify the resource internally: it becomes the label through which a user encounters that relationship elsewhere in the repository. A short title such as "Performance 4" may be convenient for data management but communicates little to a user browsing a Concert, while a more descriptive title improves context at the cost of considerable length. The current titles represent a pragmatic compromise and remain provisional as the interface develops.

The model also separates descriptive granularity from media granularity. SACM frequently retains a complete concert as a single audiovisual file, but this does not require the concert to remain the smallest descriptive unit. Individual performances can

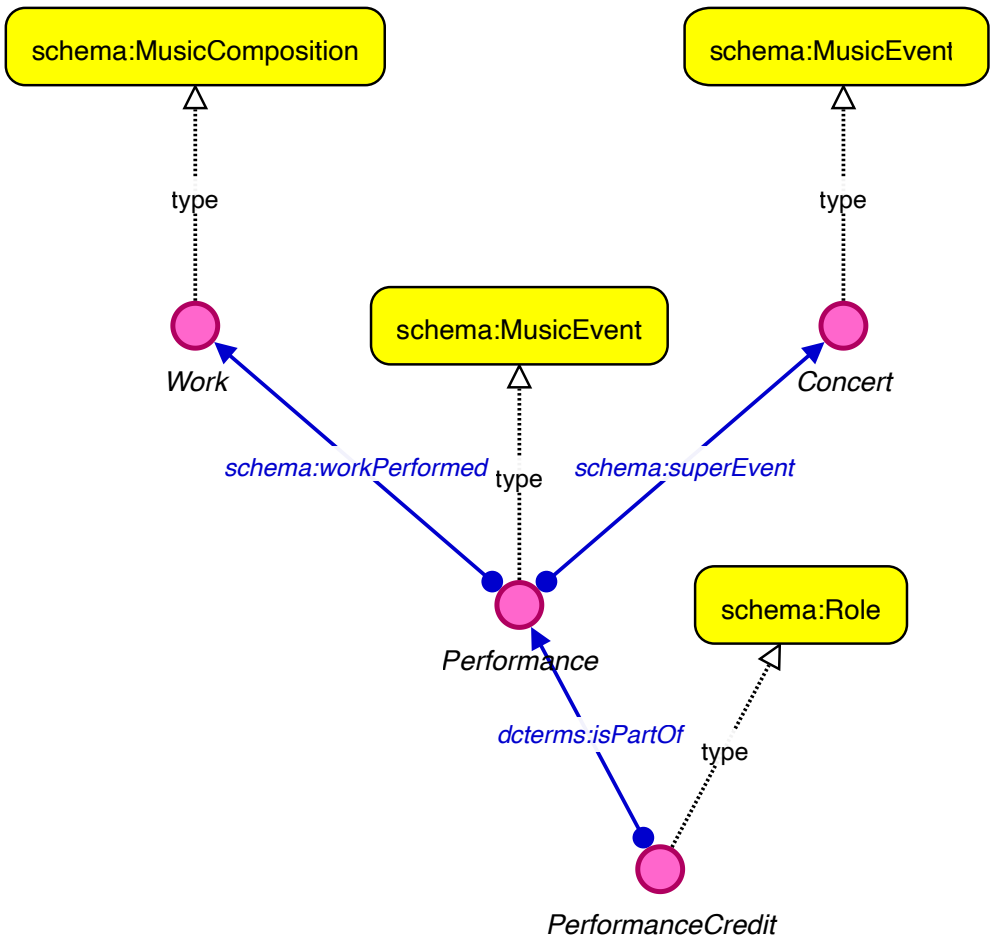

Figure 1: Conceptual data model for the SACM Recording Repository, showing the principal relationships between Work, Performance, Concert and Performance Credit. A Performance links a musical Work to a Concert, while Performance Credit represents a performer's participation in a particular Performance. Pink nodes represent repository resources, yellow boxes their classes, solid blue arrows relationships between resources, and dotted arrows class assignments.

be represented even when separate media files do not yet exist. This makes the contents of a larger recording discoverable through metadata while allowing individual performance files to be associated with those records in the future. This means that the unit through which a user discovers the collection does not need to be the same as the unit in which the media is stored.

Even where the audiovisual media remains attached to a complete Concert, the model creates potential routes to the recording through the individual Works and Performances described within it. Structured metadata could support browsing or faceting by composer, performer, work, instrument or genre, depending on the completeness and consistency of the available metadata. A user interested in flute repertoire, for example, could potentially identify Performances involving the flute and follow those relationships back to the Concert recordings in which they occur. Similarly, repeated Works or composers could be surfaced across different Concerts, allowing patterns within SACM's performance history to be explored. These discovery possibilities arise from the structure of the model and have not yet been evaluated through user testing.

These relationships can be seen in the resulting repository. The Cape Town Woodwind Festival Opening Concert is represented as a Concert resource containing a series of linked Performances. The performance of Martin Watt's Sonata No. 2 for Flute and Piano on 8 April 2025 is represented separately, linking to that Concert through `schema:superEvent` and to an independent Work

Resource	Class	Principal relationships / properties
Concert	`schema:MusicEvent`	date, location, genre, organiser, identifier; linked from Performance through `schema:superEvent`
Performance	`schema:MusicEvent`	`schema:workPerformed` → Work; `schema:superEvent` → Concert; order within concert
Work	`schema:MusicComposition`	title, composer, contributor, country of origin, identifier; linked from Performance through `schema:workPerformed`
Performance Credit	`schema:Role`	`dcterms:isPartOf` → Performance; performer, instrument, role, character, identifier

Table 1: Principal SACM resource types and their implementation in Omeka S.

Discovery term	Potential route to media
Work	Work → Performance → Concert → recording
Composer	composer value → Work → Performance → Concert
Performer	performer value → Performance Credit → Performance → Concert
Instrument	instrument value → Performance Credit → Performance → Concert
Genre	genre value → Work / Concert → recording

Table 2: Examples of potential discovery routes enabled by the SACM metadata and linked-resource model.

through `schema:workPerformed`. The Work itself can therefore connect performances across different events: Watt's Sonata No. 2 currently links to both the April Woodwind Festival performance and a later performance during the Homegrown - South African Music Showcase.

Although the resulting structure is relatively simple, two distinctions became particularly important during its development: separating a Work from each individual Performance of it, and determining how the participation of performers should be represented. The following section examines these modelling decisions in more detail.

3 Modelling Performance

3.1 Work and Performance

One of the most significant decisions in developing the SACM model was to distinguish a musical Work from an individual Performance of that work. In the original spreadsheet these could appear together within a single row: a title and composer alongside the performers and concert in which the piece appeared. For a programme this is entirely logical, but for the digital collection it combines information about a composition with information about a particular occasion on which that composition was performed. This movement from domain-specific knowledge into a formal representation of entities and relationships is a central concern of data modelling in Digital Humanities (Flanders and Jannidis, 2015). McCarty similarly describes modelling as a process through which humanities knowledge is made explicit, tested and revised through formal representation (McCarty, 2004).

The distinction became important as the data was compared across concerts. The same composition could recur at different events, performed by different musicians and in different contexts. Creating a new Work record for every occurrence would duplicate information about the composition and obscure those recurrences. Conversely, attaching concert and performer information directly to the Work would imply that information specific to one performance belonged to the composition itself.

The separation between a musical Work and its individual performances is already well established in bibliographic and music information models. The Music Ontology, for example, separates the abstract MusicalWork from a particular Performance event (Raimond et al., 2007). IFLA LRM makes a related distinction between Work and Expression, treating a specific musical performance as one form through which a Work may be realised (Riva et al., 2017). More specialised music initiatives such as DOREMUS (Achichi et al., 2018) and MusicBrainz[5] similarly distinguish musical works from particular realisations of them. In the SACM model, a Performance links to the Work through `schema:workPerformed` and separately to the Concert through `schema:superEvent`. The Performance therefore becomes the point at which the Work and the particular concert context meet.

Martin Watt's Sonata No. 2 for Flute and Piano provides a useful example. The composition exists once in the repository as a `schema:MusicComposition`, with its own identifier and composer. That Work is currently linked to two different performances: its performance during the Cape Town Woodwind Festival Opening

[5] See the MusicBrainz Work documentation: https://musicbrainz.org/doc/Work.

Concert on 8 April 2025 and another during the Homegrown - South African Music Showcase later that year. The April Performance, meanwhile, links both to the Watt Work and to the concert in which it took place.

Separating these entities therefore does more than avoid duplicate records. It allows recurrence to become visible within the collection. A user can potentially move from a Work to its different performances, while a Performance can be explored in relation to both its composition and its concert context. The same structure could support questions about recurring repertoire, the representation of composers across programmes, or the different performers associated with a particular Work.

The modelling decision therefore determines which distinctions remain visible computationally. In this sense, modelling organises information while also making particular understandings of the material explicit (Ciula and Eide, 2017). A flatter structure could still reproduce the programme information, but it would require recurrence and variation to be inferred from repeated text rather than represented through explicit relationships.

3.2 Performance Credit

Representing performers introduced a related problem. The original spreadsheet commonly stored one or more performer names together with information such as instrument, voice type, ensemble or character. A simple approach would have been to attach multiple performer values directly to each Performance. This would preserve who participated, but not necessarily the information associated with each person's participation.

The April performance of Watt's Sonata No. 2, for example, involved Liesl Stoltz on flute and Esthea Kruger on piano. The resulting Performance record therefore needs to represent both musicians and preserve which instrument belonged to each performer. Across the wider dataset the problem becomes more complex, with performer information including operatic characters, ensembles, soloist roles, voice types and different instruments.

Maintaining separate repeatable lists of performers and roles or instruments would make their interpretation dependent on the order of values remaining aligned. More importantly, it would treat the role or instrument as though it were an independent attribute, when it actually describes a particular person's participation in a particular Performance.

A comparable problem is addressed in the Music Ontology through the decomposition of complex performance events, allowing the participation of individual performers and their instruments to be represented without ambiguity (Raimond et al., 2007). The model therefore introduced Performance Credit as a separate resource. A credit connects to one Performance while retaining information specific to that participation, including the performer or ensemble and, where relevant, instrument, voice type, character or role. The relationship between performer and Performance thus becomes something that can itself be described. This resembles the problem of n-ary relations in RDF, where additional information belongs to the relationship between entities rather than to either entity independently (Noy and Rector, 2006).

This adds complexity. A Performance involving several musicians produces several Performance Credit records where a flatter structure would require only repeated performer values. The additional structure was retained because it preserved the connection between information already present in the programmes and the participant to whom it applied.

The Performance Credit also illustrates a broader principle that may be useful beyond this collection: when a relationship itself carries meaningful information, that relationship may need to become an object of description. In the SACM data, the relationship is between performer and Performance; in other humanities datasets comparable problems may arise around the roles through which people participate in events, organisations, productions or texts.

Neither the separation of Work and Performance nor the introduction of Performance Credit was fully determined before the data was transformed. Both developed through repeated examination of the programme data and attempts to test simpler structures against it. Generative AI formed part of this process, proposing possible representations and helping translate emerging decisions into computational operations. The next section considers how those exchanges contributed to the model, including moments where proposed simplifications had to be revised in response to the collection itself.

4 AI-Assisted Modelling and Methodological Provenance

The SACM model developed iteratively as the data transformation progressed. As the spreadsheet was

cleaned and restructured, questions about how particular elements should be represented emerged alongside practical questions about implementing those decisions in Omeka S. ChatGPT formed part of this process, providing a conversational space in which possible structures could be proposed, tested and revised, as well as generating Python scripts used to restructure and clean the data.

Generative AI was used here to work through modelling problems and develop computational methods for restructuring the data. The programme data remained the primary source, while proposed representations were assessed against the material and the requirements of the repository. Although their study focuses on qualitative analysis, Yan et al. similarly describe generative AI-assisted research as an iterative process shaped by human feedback, contextual judgement and validation (Yan et al., 2024).

The development of Performance Credit provides a particularly useful example. A simpler structure in which multiple performers were attached directly to a Performance was considered during the modelling process. This would have reduced both the number of records and the complexity of the eventual Omeka S import. The problem became clear when this structure was tested against examples such as the April performance of Martin Watt's *Sonata No. 2 for Flute and Piano*, which involved Liesl Stoltz on flute and Esthea Kruger on piano. Storing the performer names and instruments as separate repeatable values would preserve both pieces of information, but would not preserve which instrument belonged to which performer. I challenged the simpler representation on the basis that the instrument described a particular person's participation in the Performance rather than the Performance independently. The model was subsequently revised around Performance Credit as a separate resource.

This exchange was significant because evaluating the AI-generated proposal against the programme data required me to articulate an assumption that had previously remained implicit. The question became whether information such as instrument or character belonged to a person, a performance, or the relationship between them. Rejecting the simpler structure therefore contributed to articulating the modelling principle discussed above: where a relationship carries information of its own, that relationship may need to be described separately.

Other decisions involved similar negotiation between conceptual possibilities and practical constraints. A more extensive authority structure based on separate Person records, for example, would have created stronger links between performers appearing across the repository. It would also have significantly increased the number of resources requiring creation and maintenance. The project retained structured performer information within Performance Credit while leaving the possibility of person authorities for future development. Performer and composer names were cleaned and normalised using OpenRefine[6], but were not reconciled against external authority records or represented as separate Person resources. This was considered sufficient for the current stage of the repository, while recognising that future reconciliation with external authority sources such as VIAF, LCNAF, Wikidata or MusicBrainz could provide a stronger layer of consistency and linkage as the collection grows. The resulting model therefore reflects both the structure of the programme data and decisions about the level of complexity that could reasonably be sustained within the project. These decisions also reflected the need to consider whether the resulting structure could eventually support both the retrospective processing of older SACM programmes and the consistent addition of new material.

These interactions show how generative AI can make possible representations available for interrogation during humanities data modelling. Proposed structures can be compared with examples from the source data, modified as exceptions appear, or rejected when they fail to capture distinctions that matter to the collection. In this process, an inadequate suggestion can still be productive: explaining why it does not work may require assumptions that were previously implicit to be articulated more clearly.

The final repository, however, preserves the outcome of these decisions more readily than the reasoning behind them. This is a broader problem in Digital Humanities, where documentation has been identified as important for explaining technical structures and communicating the decisions, limitations and knowledge produced during the development of digital resources (Middle, 2023). A future user can see that Performance and Performance Credit exist as separate resources, but the

[6] https://openrefine.org/

model itself does not explain why a simpler performer field was considered insufficient, what alternatives were explored, or which practical constraints influenced the final structure.

The concept of paradata offers a useful parallel, describing information about the interpretive process through which digital representations are produced. Baker argues that paradata can capture the decisions, reasoning and selection processes that accompany the transformation of source material into a digital research outcome (Baker, 2012). Although developed in relation to computer-based visualisation, this concern is relevant to data modelling: the resulting schema records the structure that was chosen, while the reasoning through which that structure was reached may remain invisible.

Recent research within Digital Humanities also points to uncertainty around how generative AI use should be disclosed, despite broad recognition that such disclosure is important (Ma et al., 2025). Disclosure alone, however, may not capture how an AI-assisted exchange contributed to a particular methodological decision. In the SACM project, parts of this reasoning remain distributed across conversational AI exchanges, intermediate spreadsheets, generated scripts and the researcher's memory of the process.

Not every prompt and response need be preserved. AI-assisted DH projects could therefore document significant modelling decision points as part of their methodological provenance. In this context, significant decisions are those that alter the entities or relationships represented in the model, affect how the collection can be searched or interpreted, or depend on curatorial judgement that would be difficult to reconstruct from the final schema alone. Existing provenance models such as W3C PROV provide mechanisms for describing the entities, activities and agents involved in the production and derivation of digital resources (Moreau and Missier, 2013). The methodological provenance considered here extends this concern to the reasoning behind consequential modelling decisions. Such a record might identify the problem being addressed, the alternatives considered, examples from the source data used to evaluate them, the eventual decision and its effect on the resulting model. In the Performance Credit example above, this could record the loss of performer-instrument association as the problem, direct performer attachment and a separate credit resource as the alternatives considered, the Watt performance as the example used to test them, and Performance Credit as the resulting modelling decision. Where an AI system contributed a proposal or computational solution, that contribution could also be recorded without treating the system as the authority responsible for the final interpretation.

For a methodological commons concerned with reusable datasets, schemas and workflows, such documentation would help make data models more legible to those who later encounter or adapt them. A schema can show that a relationship exists; methodological provenance can help explain why it was considered necessary.

5 Conclusion

The SACM Recording Repository began as a practical attempt to make an existing collection of concert recordings and programme data more accessible and discoverable. Doing so required more than cleaning a spreadsheet for import into Omeka S. It required decisions about what the data represented and which relationships needed to remain visible within the digital collection.

Two decisions were particularly important. Separating Work from Performance allowed the same composition to be connected across different concert contexts without collapsing those performances into a single record. Introducing Performance Credit kept information such as instrument, role or character connected to a particular performer's participation. These choices shaped the structure of the collection and created potential routes for discovering works, composers, performers, instruments and individual performances, even where the associated audiovisual material remains stored as a complete concert recording.

The SACM case shows how data modelling can become part of Digital Humanities interpretation. The structure of the model determines which distinctions and relationships can be represented computationally and which connections can later be explored. Generative AI contributed to this process by helping explore possible structures, translate decisions into computational workflows and, at times, expose assumptions when proposed simplifications did not adequately represent the collection.

This also raises a question of methodological provenance. The final schema records the model that was chosen, but does not necessarily preserve the alternatives considered or the reasoning through which particular decisions were reached. If data

models are to form part of a methodological commons, their reuse may depend on documenting some of the consequential decisions made during their development alongside the technical structure itself. Making the model visible means making visible both the relationships represented within the data and enough of the reasoning through which those relationships were constructed.

Limitations

This paper examines a single institutional project and does not propose the SACM structure as a general model for music or performance collections. The model was developed in response to the characteristics of the available programme data, the discovery requirements of the repository and the practical constraints of implementing and maintaining the collection in Omeka S.

The discussion of generative AI is based on a retrospective, practice-based account of the modelling process and does not include a controlled comparison of AI-assisted and non-AI-assisted methods. The researcher was both responsible for developing the repository and a participant in the AI interactions discussed. Although generated scripts and intermediate datasets preserve parts of the transformation process, the AI exchanges were not documented systematically enough to reproduce every modelling decision from a fixed model version and prompt history. The transformation scripts and intermediate project files have been retained as part of the working documentation for the repository, but have not yet been deposited in a public research repository.

Although SACM's collections extend across a much longer institutional history, the model described here has so far been tested only against the 2025 evening-concert programme data. Applying it retrospectively across several decades of programmes and recordings will require further work on data cleaning, authority control, identifier assignment and batch processing. The local identifiers used in the pilot also encode dates and sequence numbers from the current programme structure. They work effectively as linking keys for the present workflow, but their suitability will need to be reassessed when the model is applied to older material, where dates or programme structures may be less consistent. Composer and performer names have so far been normalised using OpenRefine but not reconciled against external authority records; linking these names to established authorities would provide an important future layer of consistency and reuse.

A separate challenge concerns sustainability going forward. The pilot demonstrates that the model can be implemented through Omeka S resource templates and batch imports, but an ongoing repository will also require a consistent workflow through which new concerts, performances, works and credits can be added without repeatedly reconstructing the import process. Developing a sustainable workflow within Omeka S forms an important next stage of the project.

Finally, the repository remains in development. Questions concerning copyright, performer consent, storage, preservation and public access continue to shape the wider project, while some of the more granular media structures anticipated by the model have not yet been implemented.

Ethics

The wider ethical and legal questions surrounding the SACM Recording Repository extend beyond the metadata model considered in this paper. Much of the descriptive metadata used in the pilot, including performer names, works, instruments and concert dates, is drawn directly from programmes produced for public performances. The repository therefore reuses information that was already published in connection with those public events. This gives the programme metadata a different privacy context from the recordings themselves, while its reuse still needs to sit within UCT's wider responsibilities for the handling of personal information. Any public deployment of named performer metadata would therefore also need to establish an appropriate basis for processing within UCT's POPIA governance framework.

SACM operates within the University of Cape Town, and decisions about recordings involving students and other performers will need to take account of the University's policies and the College's own practices around recording, consent and the use of performance material. Consent to being recorded also needs to be distinguished from decisions about how a recording may subsequently be disseminated. Public access is only one possible approach: the repository could, for example, provide restricted or institutional access to some material.

The structure developed here also creates the

possibility of more granular access. Describing individual Performances within a Concert makes it possible to locate and eventually provide access to a specific performance within a longer recording. While this improves discovery, it may also make individual performances easier to extract and reuse, including as musical samples. Questions of copyright, performers' rights, licensing and permitted reuse therefore remain important considerations for the wider repository project. These issues require separate institutional and legal decisions and fall outside the present exercise, which focuses on developing a workable model for structuring programme and performance metadata.

Generative AI was used to assist with data modelling and transformation and was not used to infer or generate information about individual performers. The modelling process nevertheless involves decisions about how performers, works and their relationships are represented within the repository. While the model aims to preserve distinctions present in the source material, these descriptive choices shape how people and performances become visible and discoverable within the resulting collection. The model should therefore be understood as a practical representation developed for this pilot, with its own choices and limitations in how the underlying performances are described.

Acknowledgements

The author would like to thank the staff of the South African College of Music for initiating this project with UCT Libraries, providing the programme data used in the pilot, and generously answering questions and clarifying the metadata and performance context throughout the modelling process. The author would also like to acknowledge Theo Herbst and Shaun Karssen for their earlier work making a selection of SACM recordings available online.

References

Manel Achichi, Pasquale Lisena, Konstantin Todorov, Raphaël Troncy, and Jean Delahousse. 2018. DOREMUS: A graph of linked musical works. In *The Semantic Web – ISWC 2018*, pages 3–19. Springer.

Drew Baker. 2012. Defining paradata in heritage visualization. In Anna Bentkowska-Kafel, Hugh Denard, and Drew Baker, editors, *Paradata and Transparency in Virtual Heritage*, pages 163–176. Ashgate Publishing Limited, Farnham.

Arianna Ciula and Øyvind Eide. 2017. Modelling in digital humanities: Signs in context. *Digital Scholarship in the Humanities*, 32(suppl$_1$) : $i33 - -i46$.

Julia Flanders and Fotis Jannidis. 2015. Data modeling. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A New Companion to Digital Humanities*, pages 229–237. John Wiley & Sons.

Rongqian Ma, Xuhan Zhang, and Adrian Wisnicki. 2025. Disclosing generative AI use in digital humanities research. *Proceedings of the Association for Information Science and Technology*, 62(1):1572–1574.

Willard McCarty. 2004. Modeling: A study in words and meanings. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, pages 254–270. Blackwell.

Sarah Middle. 2023. A documentation checklist for (linked) humanities data. *International Journal of Digital Humanities*, 5:353–371.

Luc Moreau and Paolo Missier. 2013. PROV-DM: The prov data model. W3C recommendation, World Wide Web Consortium.

Natasha Noy and Alan Rector. 2006. Defining n-ary relations on the semantic web. W3C working group note, World Wide Web Consortium.

Yves Raimond, Samer A. Abdallah, Mark B. Sandler, and Frederick Giasson. 2007. The music ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 417–422, Vienna, Austria. ISMIR.

Pat Riva, Patrick Le Bœuf, and Maja Žumer. 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. International Federation of Library Associations and Institutions.

Lixiang Yan, Vanessa Echeverria, Gloria Fernandez Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI collaboration in thematic analysis using ChatGPT: A user study and design recommendations. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.