# On the description of process in digital scholarship

David De Roure[1], Pip Willcox[2], and Alfie Abdul-Rahman[1]

[1] Oxford e-Research Centre, University of Oxford, UK
{david.deroure,alfie.abdulrahman}@oerc.ox.ac.uk,
[2] Centre for Digital Scholarship, University of Oxford, UK
pip.willcox@bodleian.ox.ac.uk

**Abstract.** The techniques and tools of linked data are being successfully applied in scholarship across many disciplines. In humanities the focus has often been on describing the content of collections and expressing datasets so that they can be linked, thus improving discovery and facilitating research. This paper suggests that descriptions of processes are also useful, be they historical processes or the process of scholarship itself, and therefore worthy of attention at the intersection of semantic web and digital scholarship. We explore this through an exercise in describing the provenance of a 'born digital' and a historical artefact.

**Keywords:** algorithmic composition, First Folio, Ada Lovelace, provenance, William Shakespeare, visualization

## 1 Introduction

Scholars are often interested in establishing where something has come from, that is, the history of an artefact, be it analogue or digital. This knowledge facilitates interpretation and trust, and describing it digitally enables the use of digital tooling to visualize, search, link, and reuse this information, and thus facilitate the scholarly process.

There have been efforts to provide means of description of processes and of provenance. For example, W3C PROV provides a data model for provenance information, with multiple serializations including RDF [1]. Another is CIDOC-CRM, which provides the CRM Digital ontology and RDF Schema to encode metadata about the steps and methods of producing digitization products [2]. For our illustrative exercise we experiment with PROV, especially as it is attracting activity in multiple disciplines. We do not address here the collection, sharing, and linkage of multiple process descriptions, but as proof of concept we note previous work in RDF descriptions of workflows [3].

In the next section we demonstrate provenance representation for algorithmically generated music, arising from research into the life of Ada Lovelace. Our second case study is based on the Bodleian First Folio of Shakespeare's plays [4], a physical artefact with a more recent digital manifestation. These provide examples of what can be usefully captured in one representation, and what we

would like to be able to represent. In particular we offer the First Folio as a hybrid physical-digital case study for future work.

## 2 Numbers Into Notes

December 2015 saw the 200th anniversary of the birth of Ada Lovelace. A major symposium was held to mark the occasion, including the discussion of a thought experiment: had Ada Lovelace lived longer, and had Charles Babbage successfully built the analytical engine, what might have happened to pursue Lovelace's observation that "the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent" (note A in [5]). We called this exercise *Numbers into Notes*.

As part of this we developed an interactive tool for people to generate music from integer sequences. The workflow of the tool mirrors our hypothesized workflow involving the analytical engine: the machine runs a parameterized program to generate a number sequence, and parts of this sequence are then given to different instruments. Inspired by the use of punched cards in the Jacquard loom and the proposed analytical engine, we generate virtual 'piano rolls'. The programmer and operator (or 'attendant') were not allowed to change the numbers generated by the machine, but had full control of the mapping from numbers to notes and then from notes to instruments. The interactive tool, a single page web application (http://demeter.oerc.ox.ac.uk/NumbersIntoNotes/), provides several algorithms which illustrate the mathematics of the early 19th century (the primary example involves generalized Fibonacci sequences, reduced by modular arithmetic to produce periodic sequences).

The final stage of the workflow is to export the musical fragment in various formats, one of which is metadata with an automatically generated natural language description of the algorithm parameters, mapping, and selection. We did this to enable someone at a later stage to be able to understand how the fragment was generated or indeed to regenerate the fragment using different tooling, i.e. to reproduce the results of the experiment. For this same reason, one of the output formats is W3C PROV-N, from which RDF can be generated, as well as an SVG visualization as shown in Figure 1. These conversions use the ProvToolbox software (http://lucmoreau.github.io/ProvToolbox/).

## 3 The Bodleian First Folio

The Bodleian First Folio of Shakespeare's plays is a physical book with a digital manifestation which facilitates scholarship. First we describe in narrative form the provenance that we wish to represent.

Shakespeare wrote, or co-wrote as recent scholarship suggests, plays for his friends and fellow actors, and seems to have crafted parts to suit their particular talents. The plays, in the forms in which they reach us, are generally longer than could practically have been performed in contemporary theatres. It may not be unreasonable to suggest performances were cut according to anticipated

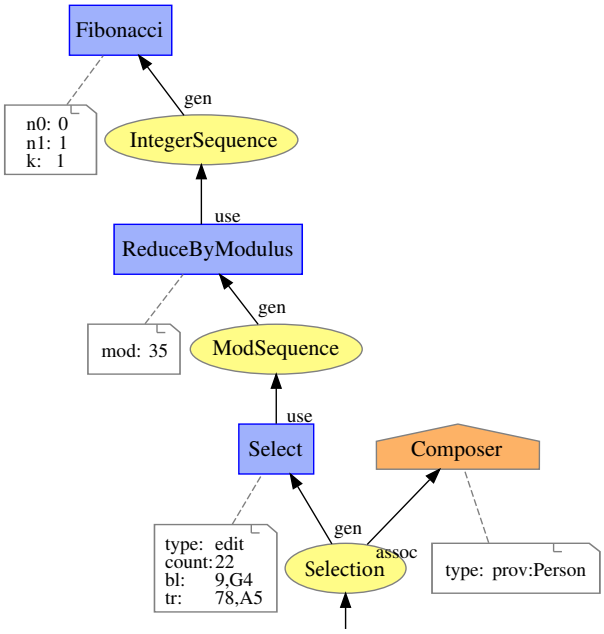


**Fig. 1.** The first part of a provenance graph generated by the *Numbers Into Notes* tool.

audiences' preferences, and not unfeasible that versions of a play were co-created dynamically, by actors responding to audience reaction.

The first collection of Shakespeare's plays (1623) is likely to derive at least in part from prompt books of the King's Men (as the company to which Shakespeare belonged was called from 1603). The First Folio, as it came to be known, was published as a joint venture by a consortium of printers – Edward Blount, William and later Isaac Jaggard, William Aspley, and John Smethwick – and two of Shakespeare's fellow actors and friends, John Heminge and Henry Condell. It republished 18 of its 36 plays with varying degrees of textual variance, publishing the other 18 for the first time.

One copy of the First Folio's print-run (estimated at between 750 and 800 copies) was sent to the Bodleian Library in Oxford, presumably under the 1610 agreement with the Stationers' Company. As was common, it arrived unbound, 'in sheets', and it was sent to a local bookbinder, William Wildgoose, to be bound strongly but plainly in brown calfskin. The book remained accessible – chained on shelf – in the Library for at least the next 40 years, apparently much read.

This copy left the library, probably sold after it had been superseded by the Third Folio of 1663/4. Lost to view for about 240 years, in 1905 Gladwyn Turbutt, an Oxford undergraduate, brought his family copy to the Bodleian Li-

brary's enquiry desk for advice on its dilapidated and lacklustre binding. The desire to return the book to its original owners inspired a private then a public funding campaign – "Oxford men" (at whom the campaign was directed, although by neither education nor gender were the donors so restricted) contributing to the local and national commons. The successful campaign saw the book returned to the Bodleian Library, still (and apparently uniquely among First Folios) in its original binding, and made fragile through frequent use by its early readers. Its physical condition meant access to the book was restricted, and few scholars were able to study it. 2012 saw a second public campaign to fund its stabilization, digitization, and publication freely online (http://firstfolio.bodleian.ox.ac.uk/).

We are developing descriptions of the provenance of this First Folio. Figure 2 shows a fragment of a simplified PROV-N visualization. It begins with the physical manifestation of a play, written out in parts for the respective actors, and as a prompt book. The co-creational activities of rehearsal and backstage annotation of the prompt book produce a text which is performed. For the sake of this exercise the text which arises from this performance is imagined as a fair copy. By the hands of Shakespeare's friends John Heminge and Henry Condell this fair copy is taken to the consortium which prints multiple copies, one of which is sent to the Bodleian Library under the agreement with the Stationers' Company. The Library then commissions its binding by William Wildgoose.

A more elaborate graph makes use of the notions of actors and plans, as with the 'composer' in our first example, and we can use the notions of *specialization* (specializationOf) and *invalidation* (wasInvalidatedBy) to capture modification and relocation of the physical work. Graphs are available on the lead author's website http://www.oerc.ox.ac.uk/people/dder.

## 4 Discussion

The PROV output from *Numbers into Notes* was definitive and achieves the desired purpose. However, working the First Folio through using PROV raises many questions. Our hypothetical premise was the usefulness to scholars of linking across provenances to the various entities and agents we declared: the John Heminge who ostensibly co-edits the First Folio's text and works with the printers' consortium is also a shareholder in The Globe theatre, the company's financial manager, husband to Rebecca Knell, a beneficiary of Shakespeare's will, and so on. At this level, when linked to other relevant data, our description could be useful.

With scant contemporary record, even aspects of provenance that are generally undisputed and verified by other research can be traced to a scholar's original work. Both to credit and to attribute the scholarship an extra field seems required, one step removed from the provenance itself – indeed, the provenance of the provenance. A level of certainty would also be helpful: one scholar's claim might logically fill a gap in the provenance, but if it were unattested elsewhere it could usefully be more hazily visualized than uncontested nodes. We also note the approach of [6] which captures links to 'spots' in primary sources.

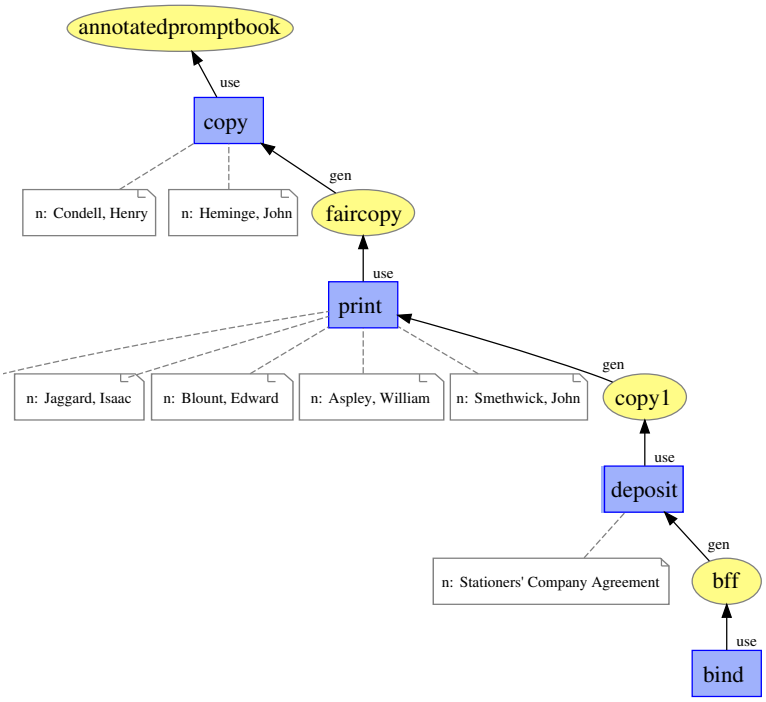


**Fig. 2.** Fragment of the provenance graph for the Bodleian First Folio.

Plurality repeatedly vexed our description. We have knowingly taken one of the world's most studied books as our subject, as a form of stress test for the encoding. While one book, it is made up of preliminaries and 36 plays, each of which has its own, much studied, history. A play's text could have more than one immediate source: a copy of a play previously printed in quarto format may have been available as its text was composed in a print shop. Print is plural by design, and variants in the text make each copy of this relatively common early modern book unique, but we describe the First Folio's provenance up to the point that a copy is sent to the Bodleian Library as though it were one work. We collapse the many processes of book production and so obscure the elements of production that make copies unique.

However, the digital phase of the book's existence yields well to PROV: scans of pages, transcription to TEI (including annotations based on the materiality of the physical book), renderings on screen, and downloads in XML and PDF. We offer the Bodleian First Folio as a challenge in process description encompassing the analogue and digital, involving uncertainty and itself the subject of scholarly process, and hope that others might encode it comprehensively in current or future representations.

The exercise of capturing the provenance has proven interesting in its own right, using a mixture of drawing and hand-encoding, and we suggest that interactive visualizations of provenance are useful tools. This is further evidenced by visualization work in humanities, such as visual analytics for intertextuality [7] and poetry visualization [8]. The algorithmic composition example captures the provenance of a sonification, which is essentially a kind of visualization, and we suggest that describing the provenance of a visualization has similar utility in interpretation and reuse of scholarship.

**Acknowledgements** This work is partially supported by *Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption* funded under EPSRC grant EP/L019981/1. We are grateful to Graham Klyne for his advice on using PROV.

## References

1. Gil, Y., Miles, S., Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., Zednik, S.: PROV Model Primer. Working group note, W3C (2013)
2. Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., Melessanakis, V.: Modeling and querying provenance by extending CIDOC CRM. Distributed and Parallel Databases, 27(2), 169–210 (2010)
3. Newman, D., Bechhofer, S., De Roure, D.: myExperiment: An ontology for e-Research. In Semantic Web Applications in Scientific Discourse, volume 523 of CEUR Workshop Proceedings: http://ceur-ws.org/Vol-523/ (2009)
4. Shakespeare, W., Heminge, J., Condell, H., Droeshout, M., Jaggard, I., Blount, E., Jaggard, W., Smethwicke, J., Aspley, W.: Mr. William Shakespeares comedies, histories, & tragedies. Published according to the true originall copies. Printed by Isaac Iaggard, and Ed. Blount at the charges of W. Iaggard, Ed. Blount, I. Smithweeke, and W. Aspley, London. Oxford, Bodleian Library, Arch. G c.7 (1623)
5. Lovelace, A.A.: Sketch of the analytical engine invented by Charles Babbage, with notes by the translator. In Scientific Memoirs, Selected from the Transactions of Foreign Academies of Science and Learned Societies, Vol. 3, 1843, pp. 666-731, volume 3. Richard and John E. Taylor, Red Lion Street, Fleet Street, London. Translation of *Notions sur la machine analytique de M. Charles Babbage* by Luigi Federico Menabrea, in Bibliothèque Universelle de Genève, nouvelle série 41, 352–76 (1842)
6. Pasin, M., Bradley, J.: Factoid-based prosopography and computer ontologies: towards an integrated approach. Digital Scholarship in the Humanities, 30(1), 86–97 (2015)
7. A. Abdul-Rahman, A., Roe, G., Olsen, M., Gladstone, C., Whaling, R., Cronk, N., Morrissey, R., Chen, M.: Constructive visual analytics for text similarity detection. Computer Graphics Forum, doi:10.1111/cgf.12798 (2016)
8. Abdul-Rahman, A., Lein, A.J., Coles, K., Maguire, E., Meyer, M., Wynne, M., Johnson, C.R., Trefethen, A., Chen, M.: Rule-based visual mappings – with a case study on poetry visualization. Computer Graphics Forum, 32(3), 381–390 (2013)