the yellow nineties online
edited by Dennis Denisoff and Lorraine Janzen Kooistra

From TEI to Linked Open Data: Crossing the Stile

By Alison Hedley (Team lead of *The Yellow Nineties Personography*)

A version of this paper was presented at the annual conference of the Canadian Society for Digital Humanities/Société canadienne des humanités numériques in May 2017. The paper was part of a panel on which members of the Ryerson Centre for Digital Humanities addressed the topic of futures of biographical encoding.

Peter Gavigan, "Stile du chemin de la messe, du champ, Ranaghan." *Wikimedia*.

In Fall 2013, I became the team lead for the *Yellow Nineties Personography*, a biographical database of the persons who contributed to the aesthetic periodicals remediated in digital editions on *The Yellow Nineties Online*. The goal of the personography project, as outlined by its primary investigator, Lorraine Janzen Kooistra, is to understand the social networks that produced these magazines. Increasing the visibility of lesser-known periodical contributors, including marginalized groups such as women and colonial subjects, has become a major facet of this work.

In our initial vision for the project, we planned to develop a personographic dataset in a spreadsheet and transform it into a TEI-encoded document. TEI (Text Encoding Initiative) is an encoding system, compliant with XML (eXtensible Markup Language), widely used by digital humanities scholars.

We anticipated several outputs from the project, foremost among which would be a user interface for exploring the data and learning about the project's methodology. Should our interface prove to be limiting for other researchers, the XML files of the personography would be available on the website for anyone to download and use. We also planned to publish findings that emerged from our analyses of the personographic data.

Between 2013 and 2016, our team trawled digital and paper archives, encyclopaedias, biographies, databases, census documents, digitized periodicals, and publication records, gradually accumulating information about the 351 individuals in our spreadsheet and their relationships with one another. We revised our spreadsheet's data fields and developed controlled vocabularies as we went along. Toward the conclusion of this initial research phase, I attended a course at the Digital Humanities Summer Institute (DHSI) that initiated me in the process of writing extensible stylesheet language transformations—which would transform our personography from csv, the file format of spreadsheet data, into TEI and other manipulatable XML formats. At another DHSI course, I learned how to craft an ODD (One Document Does it all)—the document that would formally describe how our project uses TEI.

We had not advanced far into this second phase of personographic development when, in December 2016, a question emerged from the discussion at a meeting of all members of the Ryerson Centre for Digital Humanities: was TEI really the ideal primary data format for this project? We had taken for granted that the primary data format of the personography would eventually be TEI XML because the foundational scholarly dataset of *The Yellow Nineties Online*, a body of data annotating 1890s periodicals, is written in TEI. But was TEI the most appropriate encoding system for our personographic dataset? Although I had been thinking about the personography within a TEI framework for as long as I had been on the project, I immediately recognized the value of carefully considering this question. It prompted a realization that, looking back, seems to have been just under the surface of my engagement with the personography for a while. TEI was no longer a great fit for our project as we had come to envision it. The crux of the matter is this: TEI imposes a genealogical

hierarchy on data. Such a structure is an excellent model for data that is temporally and/or textually oriented, but the *Yellow Nineties Personography* is neither.

The TEI Guidelines can facilitate superior work in biographical encoding, as demonstrated by the personographies of projects such as the *Map of Early Modern London*. Indeed, TEI coined the term "personography" and its associated practices (Flanders). A typical TEI personography comprises a collated list of the persons mentioned in encoded texts; its organization foregrounds the relationship between those persons and the texts that serves as the foundational documents of a given TEI encoding project. Its primary focus is on marking up texts. Most literary texts, such poems, plays, and novels, can be approached as hierarchical documents divided into structured units of data. This is the OHCO (Orderly Hierarchy of Content Objects) model of text (Renear et al). As the TEI Guidelines point out, OHCO is a "grossly simplified view of text" but offers sufficient infrastructure for organizing textual data in TEI (Burnard and Bauman, "Gentle Introduction" v. 4).

The TEI Guidelines include protocols for documenting pieces of data that do not tidily descend in the hierarchy but overlap with one another. One might combine different types of hierarchies to show overlapping types of data organization—for example, division of text by stanza and by page (Burnard and Bauman, "A Gentle Introduction" v. 4). One might also use the "pointing" mechanisms within TEI to document elements as having attributes shaped by a logic that is external to the hierarchy in which that element is positioned in the code (Burnard and Bauman, 16.0.0). However effectively they are deployed, these protocols are workarounds; their framing as such underscores that TEI has been designed to describe data in hierarchical terms, rather than relational ones (Burnard and Bauman, *v*.2). TEI's lack of a relational orientation is also evident in the guidelines for documenting persons. The TEI Guidelines assign data about persons, places, and organizations into three categories: *traits*, which do not change over time; *states*, which are temporally specific; and *events* (Burnard and Bauman, 13.3.1). A relationship might be assigned to any of these three categories, but it is not considered sufficiently fundamental to be an equivalent

element in the module of the Guidelines for Names, Dates, People, and Places. Most TEI personographies omit relationships, emphasizing temporal or text-based characteristics of persons.

Given its hierarchical orientation, then, TEI is not conducive to modelling and interpreting relational data. As our development and understanding of the *Y90s Personography* advanced, we began to consider whether the hierarchical organization of TEI was becoming a limitation. We wished to emphasize the relational, socio-cultural roles of historical persons without organizing these into a temporal or text-based hierarchy. Our biographical data includes some information that might be modelled in such a hierarchy, such as birth and death dates and metadata about contributions to specific periodicals. However, the most innovative and extensively researched fields of our personography documented the relationships that gave shape to a social network of cultural production.

In terms of its suitability for the *Yellow Nineties Personography*, another limitation of TEI is that it is not highly interoperable in practice. TEI is designed to promote interoperability; as one of thousands of variants of XML, it can link to other datasets with varying degrees of success, depending on how a specific project or author uses it. The TEI Guidelines ensure that a set of rigorous standards facilitates the interchange of data among researchers using different programs, systems, and software (Burnard and Bauman, *iv*). However, data interoperability is fairly limited in practice. The TEI Guidelines do not require interrelation as a primary function of datasets. Even if published as open-access scholarship, many TEI-encoded texts remain isolated from other TEI datasets in that they do not reference one another or otherwise interconnect to enable users to interpret data across multiple projects. In a TEI-encoded text, XML identification numbers and the reference attribute can be used to point to data outside a specific passage of XML, but these are normally used to reference other portions of the same text and additional XML files containing derivative data such as a placeography. In TEI, any linked data remains closed within an internal ecosystem.

While neither relationality nor interoperability is central to the TEI principles, these characteristics have become increasingly central to our conceptual model for the *Yellow Nineties Personography* as the dataset has expanded and our nuanced understanding of its possibilities has deepened. As a result, the project team arrived at a turning point familiar within the broader context of the humanities research process. I think of this point in the research process as the crossing of a stile—perhaps because I'm a scholar of nineteenth-century print media and stiles frequently appear in Victorian literature.

A stile enables someone travelling on foot to cross over the fence enclosing a portion of land. It allows the individual to occupy a liminal space, briefly poised between two lands at an elevated vantage point from which they can survey their progress. The stile offers a useful metaphor for the transitional moments that every scholar confronts during the research process. An initial spark of curiosity prompts exploration, reflective analysis, and a hypothesis that informs further exploration. In light of advancing knowledge, the hypothesis evolves until a strong argument takes shape. This process often involves crossing a few stiles, moving into different territory as the scope, method, and argument change. The bigger the project, I find, the more stiles a scholar crosses.

In DH, crossing the stile is often articulated in terms of troubleshooting, iteration, and productive failure. As John Unsworth observes, "If failure isn't a possibility, neither is discovery" (4). Rigorous testing of a theory or method risks failure by seeking to verify it. In the context of humanities research and analysis, digital and otherwise, initial failure spurs revision and iteration: if one type of visualization or transformation scenario did not yield fruitful results, figure out why and try another one. Such setbacks improve our understanding of the makeup of our data and the functionality of our tools while enabling us to refine our research questions.

No longer sure that modelling our data in TEI was the most effective way to realize our desired outcomes, we stood on our stile and decided to take the personography into the terrain of linked open data (LOD). LOD involves linking information across databases that have a non-relational structure—in other words, their data strings are organized into RDF (resource description

framework) statements that declare relationships between pieces of data using a subject/predicate/object structure (also called triples). The non-relational data model is highly customizable, so it lends itself readily to structuring heterogeneous datasets. If a non-relational database takes a structure and vocabulary that are not wholly unique, but have equivalents in other non-relational databases, these data can be recognized and processed by myriad types of software once published online. LOD describes a set of best practices for structuring non-relational RDF data so that they can converse with datasets originating from different types of projects. LOD is useful for anyone who wants the data they publish online to be easily discoverable and shareable by other computers and other researchers. It has many champions in computer sciences, library and information sciences, and digital humanities, where information sharing and community brainstorming are importance aspects of scholarship. Open access is a central tenet of LOD and linked open datasets are released under an open license, so they can be reused by anyone.

Our primary goal for the project remains the same: to adumbrate the social networks that produced the periodicals of *The Yellow Nineties Online*, with an emphasis on documenting lesser-known persons in this network. However, our methods for realizing these goals are substantially different than what we first envisioned. Modelling our personography as LOD will allow us to maintain a relatively heterogeneous dataset while linking this data to other records of our persons that exist elsewhere, such as the Library of Congress or other DH projects addressing the magazine artists, authors, and editors that we do. Once we have transformed our dataset from CSV to RDF format, we will develop a user interface for interacting with that data. Before we can transform the data, however, we must reorganize the dataset's structure and vocabulary, developing a formal description that is known, in LOD practice, as an *ontology*. We have undertaken this task with three considerations in mind. The first is how we can most effectively organize our data in relation to one another. The second is the common ground between our entity and property types and those developed for other LOD ontologies, using previously established terms as much as possible. The final and most practical consideration in developing our ontology is how we

can ensure that the personographic data that is currently in spreadsheet form will be coherent once transformed into RDF. This is less conceptually challenging than it might seem. In fact, we had already started organizing some of our data fields into subject/predicate/object strings that resemble RDF statements—for example, "H. G. Wells (subject) is friends with (predicate) E. Nesbit (object)."

Interestingly, switching to LOD has immediately drawn us into a community of collaborators—partly because we are in unfamiliar territory and need much guidance, and partly because linked open data necessarily involves using the data structures and best practices of others as much as possible. Our collaborators on the LOD process include three Ryerson librarians, Naomi Eichenlaub, Trina Grover, and M. J. Suhonos, as well as members of the international digital humanities community, notably the DH librarian at Miami University, Paige Morgan. We have discovered that the field of linked open digital humanities is largely uncharted, with no clearly demarcated trail to guide our steps. However, we periodically compare notes on our workflow and ontology with other projects, particularly the *Linked Modernisms Project* and the *Canadian Writing and Research Collaborator Ontology*.

In keeping with the principles of LOD, we are developing the *Yellow Nineties Personography* with an eye to its legacy. In terms of linked open scholarship on the nineteenth century, the personography's main contribution will be the recuperation of contextual information about now obscure individuals and social networks that we argue were significant to Victorian cultural production—such as the ornament designers Helen Hay and Annie Mackie and the rest of a Scottish circle of Celtic revivalists who created *The Evergreen: A Northern Seasonal* (1894-1895). Hay and Mackie, like many of the female contributors to the periodicals of the *Yellow Nineties Online*, are not well known to Victorian scholars. The *Yellow Nineties Personography* will increase their visibility by adding to pre-existing internationalized resource identifiers (IRIs) and, as necessary, creating new ones. An IRI is a unique sequence of characters that designates an object or resource. For biographical encoding, a person's IRI serves as the basis for linking records of that person that appear in datasets around the world. Information and links generated from these instances are

collated on a kind of virtual IRI index card that is published on VIAF, a database listing all virtual international authority files. In linked open biographical encoding, the IRIs of persons are essential, because they offer the most foundational type of data: documentation that signals the existence of persons and lists the archives where records about them will be found. VIAF includes multiple entries for the name Helen Hay, but not one unambiguously matches our personographic entity. Annie Mackie has no VIAF entry, and therefore no IRI. By determining which IRI best corresponds with the designer Hay and generating an IRI for Mackie, we declare these individuals as culturally significant enough to warrant linked open documentation and create the opportunity for others to make closed records of these persons part of a shared biographical network. For entities in our personography who already have IRIs, such as John Duncan, who trained Hay and Mackie in designing ornaments, linking our dataset to VIAF will increase the visibility of information about their lives and social networks that is not discoverable through major resources such as the *Oxford Dictionary of National Biograp*hy or Wikipedia.

The *Yellow Nineties Personography* makes a critical intervention in the politics of linked open biographical data. The non-relational database structure's capacity to accommodate both customization and interoperability means that modelling linked open data involves negotiating the irresolvable tension between homogeneity and heterogeneity. The most complete realization of the LOD principle of interoperability would involve imposing the same structure and vocabulary on all linked data. However, imposing such a structure would erase the contingencies, ambiguities, and idiosyncrasies of cultural records. Our dataset contains many anomalies that we do not wish to elide—that we wish, in fact, to celebrate. Respecting the heterogeneity of entities is crucial to the argument we want our data model to make. Fortunately, this tenet is enshrined in LOD practice through the principle of customization.

Two examples illustrate how we are working to balance heterogeneity and homogeneity in the *Yellow Nineties Personography*. First, we have developed a Victorian-specific taxonomy of the occupations our persons held over their lifetime, something we believe no other LOD ontology has yet done. Although we

will use a historically specific classification system, we will try to facilitate interoperability by using occupation terms that exist in other LOD vocabularies. Another example more clearly foregrounds what is at stake in negotiating this tension between universality and customization. VIAF displays multiple international authority files for a personography entity we have entered under the name Sarojini Chattopádhyây. Institutions around the world have recorded this entity's name in multiple ways, some listing her maiden name and some her married name. A few records of this entity include the punctuation of Romanized Bengali, which is how it appears in the *Savoy Magazine* curated on *The Yellow Nineties Online*. A handful of other entries reproduce the name in Bengali characters. However, the records that appear at the top of the search results for this person in VIAF omit any trace of Bengal. The entries that have been imbued with the most authority, by virtue of linking to the greatest number of library and museum catalogues online, document this person as "Sarojini Naidu," privileging a more wholly Anglicized variant of the name. The most internationally recognizable variant of this person's name is imperialist in its Anglophone origins, and to prioritize this variant in our database would be to reinscribe the history colonial violence. The *Yellow Nineties Personography* uses a lesser-known version of this entity's name because it is a spelling that was recognized and most in accord with her Bengal culture. We will document such choices as general practices in our ontology, so that our personography will serve as a model for others developing linked open biographical datasets with political sensitivity.

In January 2017, linked open data presented the most promising prospects for the *Yellow Nineties Personography*. Subsequently, as we have worked to implement LOD into our biographical encoding practice, we have discovered more about the possibilities and opportunities the methodology offers. I suspect that the principles of LOD will have lasting influence on biographical encoding practices in the digital humanities generally. It brings together humanities scholars, students, librarians, and other faculty across institutions. I envision this future of biographical encoding as continuing to draw on other data models, including that of the TEI. Significantly, our personography still uses TEI in a number of ways. The biographical dataset is still linked to periodicals that the

Yellow Nineties team has marked up in TEI, and the related projects use a common set of unique identifiers for magazine contributors. We hope that eventually our personography will connect to the XML data about the features of the periodicals, though this goal is further in the offing.

If linked open data gains traction in the digital humanities, and particularly in biographical practice such as personography, the future of biographical encoding will involve perennially negotiating the tension between interoperability and customization. However, all digital humanities work requires a critical balance between the disambiguating work of computation and the wonderfully messy interpretive layers of our cultural records.

Works Cited

Burnard, Lou and Syd Bauman. *P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.1.0. Text Encoding Initiative, 15 Dec. 2016, www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Flanders, Julia. "Encoding Textual Information." *Women Writers Project*, Northeastern University, 27 Sept. 2011, wwp.northeastern.edu/outreach/seminars/walden/presentations/contextual_encoding/contextual_encoding_00.xhtml.

Gavigan, Peter. "Stile du chemin de la messe, du champ, Ranaghan." *Wikimedia*, April 2007, commons.wikimedia.org/wiki/File:Stile_02_Ranaghan.jpg.

Renear, A., E. Mylonas, and D. Durand. "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies." Nancy Ide and Susan Hockey, editors. *Research in Humanities Computing*, Oxford UP, 1996.

Unsworth, John. "Documenting the Reinvention of Text: The Importance of Failure." Journal of Electronic Publishing, vol. 3, no. 2, Dec. 1997. doi:10.3998/3336451.0003.201.