BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 9 (2024) 1–12

BRILL OPEN

brill.com/rdj

Two Datasets of Genetic Dossiers and Authorial Manuscripts

Elena Spadini | ORCID: 0000-0002-4522-2833
Research and Infrastructure Support, University of Basel, Basel, Switzerland
elena.spadini@unibas.ch

Received 27 August 2023 | Revised 23 May 2024 |
Accepted 14 June 2024 | Published online 12 September 2024

Abstract

In genetic criticism, scholarly editing, authorial philology and, more generally, for the study of authorial manuscripts and writing processes, it is essential to order and classify the textual witnesses and their relationships. This article presents two datasets of so-called 'genetic networks', that is representation of the genetic entities (witnesses, publications, dossiers) and their relationships, modelled according to the GENO 1.0 ontology. The datasets contain genetic networks of the works of two Swiss authors: the main publications of Gustave Roud (1897–1976) and the short story "En mer" by Bernard Comment (1960).

Keywords

genetic criticism – textual scholarship – scholarly editing – authorial philology – digital philology

– Related data set "Genetic networks datasets: ten books by Gustave Roud, "En mer" by Bernard Comment" with DOI www.doi.org/10.5281/zenodo.8287801 in repository "Zenodo"

Published with license by Koninklijke Brill BV | DOI:10.1163/24523666-bja10042
 | ISSN: 2452-3666 (online)

1. Introduction

Genetic criticism and authorial philology are fields of study concerned with the genesis of literary, and more in general, textual works (in English, see Van Hulle, 2022; Italia & Raboni, 2021). Analysing the archival materials that document the creation process can shed new light on the dynamics of writing, as well as on the texts themselves, which is relevant for literary history, literary criticism, and textual scholarship at large. A document that gives testimony of a writing project, resulting or not in a published work, is called a witness, or genetic witness: a sketch, a fair copy, a list of names, or a document containing quotations from other works, are all examples of genetic witnesses. They can exist in different formats and on different supports: for example, paper or digital, manuscript or typescript.

A typical workflow in genetic criticism and authorial philology consists of a first phase in which the materials are collected, analysed and organised chronologically within dossiers; and a second phase, in which the documents are transcribed and edited, and a study of the genesis is carried out (De Biasi, 2008). In a digital paradigm, a suitable standard exists for the work of the second phase: the Text Encoding Initiative (TEI), used for transcribing and editing genetic documents, allows to model what happens on the page of a single witness, as well as to model the textual variants among the witnesses. This article focuses instead on the first phase of the workflow and, in particular, on the classification and ordering of the documents (in French, *classement des manuscrits*), a step that is often neither explicitly modelled nor formally documented.

The classification and organisation of the witnesses are generally implied in a genetic edition, whether print or digital: the documents are listed chronologically and presented one after the other. The edition almost always represents a single work, with all the genetic materials collected in a genetic dossier. The relationships between the documents and possibly also between dossiers are analysed in a genetic essay or in a textual note accompanying the edition. The work presented in this article complements the usual genetic edition as described here, by making explicit and formalising the classification and ordering of the witnesses in computable data that can be published as part of the edition.

The data presented here contains so-called 'genetic networks'. A genetic network is a representation of the genetic entities (witnesses, publications, dossiers) and their relationships, expressed in a formalised way (Spadini et al. 2023). For example, a genetic network might consist of all of the following elements and the links between them: three witnesses – a diary note, a draft and

a clean copy – that belong to a dossier; the dossier, that leads to a publication in a literary journal; the publication that is reused with other materials to write a book. This article documents two datasets of genetic networks. The model used to formally structure them is the GENO OWL2 ontology. GENO is a generic model for describing genetic entities and their relationships, applicable in the fields of genetic criticism, scholarly editing, authorial philology and, more generally, the study of authorial manuscripts and writing processes.

Genetic networks in the form of findable, accessible, interoperable and reusable (FAIR) data can give new impetus to genetic criticism and authorial philology. The networks allow scholars to pose questions directly to the structured data, in addition to reading the result of an expert's analysis in an essay or note to the edition. This is particularly important in an exploratory phase, for example when faced with a large amount of data, with a new case study or with a comparative analysis of more than one genetic profile. Some of the questions that can be asked to the datasets modelled with GENO are, for example, which of the published works of some authors are rooted in their diaries; how many witnesses belong to a dossier and to how many dossiers a witness can belong to; in which dossier is a particular publication reused; which documents belong to a pre-compositional phase. Datasets of genetic networks can also be useful when the texts themselves are not available due to copyright or other restrictions, a common situation in this research field (Dillen & Neyt, 2016): by creating datasets, experts can share many of the insights gained while working on a corpus without violating any rights.

This article presents the research problem that the datasets aim to address, the datasets themselves, and how they have been created and used to date.

2. Problem

Structured data that represents the classification and ordering of the genetic witnesses and of the dossiers does not yet exist. To our knowledge, GENO is the first data model to address this step of the genetic workflow, and the datasets presented here are the first of their kind.

In classical and medieval philology, the classification and ordering of the witnesses corresponds to the *recensio*, leading to the creation of a *stemma codicum*. As Van Hulle noted (2022, p. 10), the notion of stemma is seldom used in genetic criticism, and its underlying arboreal model is not suitable for representing the genesis of textual works. Visualisations similar to a stemma but conceptualised differently, as maps or tables, are sometimes employed in editions and essays to represent the relationships between genetic materials.

Examples can be found on the page "Tableaux génétiques" in *Les manuscrits de Madame Bovary* (Girard et al., 2009), "Das genetische Dossier" in *Hermann Burger, Lokalbericht. Digitale Edition* (Dängeli et al., 2016), "Stemmata" in *Friedrich Dürrenmatt. Das Stoffe-Projekt* (Probst & Weber, 2021), "Les relations entre toutes les notices du corpus" in *Robinson de Paul Valéry: édition génétique* (Johansson et al., 2018), "Manuscript Chronology" in the editions of the *Samuel Beckett Digital Manuscript Project* (Van Hulle & Nixon, 2011–2022). These diagrams convey the information visually but they are not data visualisations because they are not produced by processing underlying structured data. In contrast, the datasets presented in this article contain structured information and have been used to produce data visualisations (see section Methods below).

The lack of a data model for representing genetic networks has been addressed by the creation of GENO, the Genetic Networks Ontology (Spadini, 2023b). GENO is a generic data model that allows to represent classification and ordering of the genetic materials in RDF statements. GENO distinguishes between witnesses (an existing instance of a text, unique and characterised by the close relationship between the text and its support), publications (the realisation of a work, a subclass of the FRBR class Expression) and dossiers (the collection of witnesses that provide evidence for a writing project). Witnesses are members of dossiers, and the dossier may result in a publication, which in turn may be used in other dossiers. Witnesses can be arranged chronologically and further defined as avant-textual witnesses (belonging to a particular genetic stage), diary entries, documentation, marginalia or other types of material. Furthermore, the concepts of endogenesis and exogenesis, as well as genetic phases (pre-compositional, compositional, pre-publication, post-editorial), are operationalised in the latest version of GENO (version 1.0).

3. Data

- **Data file name deposited at source – DOI:** www.doi.org/10.5281/zenodo.8287801
- **data_GustaveRoud_allBooks_geno1.0.ttl**
  - Roud dataset
- **data_BernardComment_EnMer_geno1.0.ttl**
  - Comment dataset
- **Temporal coverage:** 1927–1972 (Roud), 1998–2011 (Comment)

This article presents two RDF datasets modelled according to GENO (version 1.0). The two datasets are:

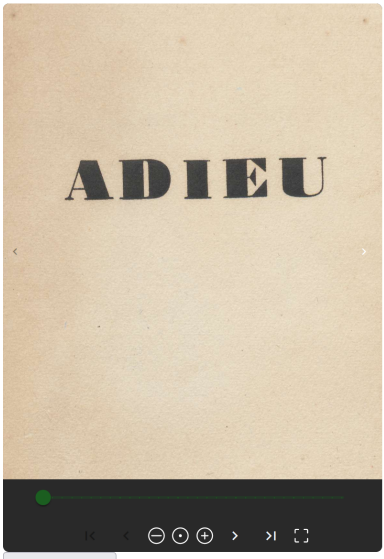
– Roud dataset: 'data_GustaveRoud_allBooks_geno1.0.ttl' contains 5680 RDF triples (here in the Turtle serialisation) that represent the genetic networks of Gustave Roud's major publications. Gustave Roud (1897–1976) was a Swiss writer, translator and photographer, whose work is particularly interesting from a genetic point of view because of the extensive reuse of his own textual material in the creation and writing process: Roud reuses passages from his diary and from previously published works to compose new texts. The relationships between the genetic materials form complex networks that go beyond the individual dossier, as it is impossible to study the genesis of a work without taking into account other works.

  The GENO data model and the dataset presented here were originally created as part of the project "Gustave Roud, *Œuvres complètes*" (https://data.snf.ch/grants/grant/157970); the complete project data is published in *Gustave Roud. Textes & Archives* (Jaquier & Maggetti, 2023). The dataset presented here includes all the genetic-related data from the project, but excludes other information that is not directly relevant to the genetic networks, such as the full text of the works and the named entities mentioned in them. In the project, genetic entities and relationships were modelled according to the previous version of the GENO ontology (version 0.2); here, the data has been ported to the newer version of the GENO ontology (version 1.0).

  The dataset contains the genetic networks of Gustave Roud's major publications, ten books published during his lifetime, between 1927 and 1972. These networks exemplify different types of genesis: to mention just a few, in *Feuillets* (1929) the author draws heavily on the diary; in *Air de la solitude* (1945), he reuses more than twenty previously published texts; there are seventy avant-textual witnesses for *Requiem* (1967), a book for which very little was reused. The dataset combines the genetic networks of the ten publications, creating a large network that links 125 publications and 440 genetic witnesses. The networks can be very complex because each section of a book usually has a separate genesis and thus a specific genetic dossier. For each of the ten books, up to two levels of reuse are taken into account, i.e. the genesis of an article reused in the book is included, but not the genesis of an article reused in another article reused in the book.

  The data is published online in the 'Archives' section of the web application *Gustave Roud. Textes & Archives* (Jaquier & Maggetti, 2023). For each one of the ten major publications, the two panels 'Genèse' and

Adieu, 1927



Référence bibliographique
Gustave Roud, *Adieu*, Lausanne, Le Verseau, 1927.

Genèse (matériaux utilisés pour l'élaboration du texte)

Notes de journal

Pour la partie : « Adieu »
- Feuillet de cahier détaché, [1916] (CLSR GR MS 4 A/2a)
- « (Notes d'Est[avayer] 24 avril [1927]) » (partie de : CLSR GR MS 4 B/10a)
- « [1926] » (partie de : CLSR GR MS 4 B/14a)
- « 29 septembre-2 octobre [1926] – Nuit à Estavayer » (partie de : CLSR GR MS 4 B/10a)

Témoins génétiques
- Adieu (Collection privée)
- Adieu (Collection privée)

Pour la partie : « [Cube de corps et de fumée] »
- Ébauche d'un poam (CLSR GR MS 1 A/1a)
- Pardonne à ce cœur dur, Bain (CLSR GR MS 1 A/1b)

Pour la partie : « Adieu »
- Incipiunt laudes quas fecit BEATUS FRANCISCUS (CLSR GR MS 1 A/3a)
- Adieu (Collection privée)

Publications antérieures

Pour la partie : « [Cube de corps et de fumée] »
- *Cube de corps et de fumée*, Imprimerie Zwahlen, 1922

Réédition de

Reprises

Réemploi du texte dans la genèse d'une nouvelle publication

Pour la partie : « [Cube de corps et de fumée] »
- « Annonce d'un Adieu », *Présence*, 1932

Rééditions
- *Adieu*, Éditions des Portes de France, 1944

FIGURE 1 Page of Roud's book "Adieu" at https://roud.unil.ch/archive

'Reprises' provide the link to all the materials in the genetic network of the work, categorised and sorted (see Figure 1).

– Comment dataset: 'data_BernardComment_EnMer_gen01.0.ttl' contains 92 RDF triples (here in the Turtle serialisation) that represent the genetic network of the short story "En mer" by Bernard Comment. The data has been produced as part of a scholarly edition prototype (Spadini, 2023a). As such, it is not meant to be comprehensive, but it is representative of a hybrid (born-digital and paper) genesis and, unlike the Roud dataset, it includes genetic dossiers that do not result in publication but are reused in new writing projects.

Bernard Comment is a Swiss writer, translator and publisher. His archives are kept at the Literary Archives of the Swiss National Library. The short story "En mer" was first conceived for the radio in 1998 and published in a revised version the following year in *La Nouvelle Revue Française*. During the first

decade of the 2000's, the author included it in various writing projects that eventually do not result in any publication. A new version of "En mer" was published in *Tout passe* (éd. Christian Bourgois, Paris, 2011), winner of the Prix Goncourt de la Nouvelle. The genetic documents include plans, sketches, drafts, clean copies and corrected proofs, both handwritten and born-digital. A diagram summarising the relationships between the publications and the genetic witnesses is available at https://comment-enmer.github.io/witnesses.html#graph.

As an example of how the datasets look, the following code snippet, taken from the Roud dataset, documents the relationships between an avant-textual witness (a clear copy) belonging to a dossier, which results in a published article, which in turn is reused in another dossier.

```
@prefix geno: <https://w3id.org/geno#>.
@prefix fabio: <http://purl.org/spar/fabio/>.

<https://ark.dasch.swiss/ark:/72163/1/0112/QKyBsUJoRyW
4JphDsWdRLQ8> a geno:Avant-textualWitness;
  geno:hasGeneticStage geno:ClearCopy;
  geno:isMemberOfDossier
<https://ark.dasch.swiss/ark:/72163/1/0112/ctZuEa
2LTL=PmBoYPWABXQY>.

<https://ark.dasch.swiss/ark:/72163/1/0112/ctZuEa2LTL=
PmBoYPWABXQY> a geno:GeneticDossier;
  geno:resultsInPublication
<https://ark.dasch.swiss/ark:/72163/1/0112/RzaFM7
DjQDGal0kKZU7UCgt>.

<https://ark.dasch.swiss/ark:/72163/1/0112/RzaFM7Dj
QDGal0kKZU7UCgt> a fabio:Expression, geno:Publication;
  geno:publicationIsReusedInDossier
<https://ark.dasch.swiss/ark:/72163/1/0112/uFwo=VxATUy
=mhxashFzbwf>.
```

GENO is quite small and simple because it focuses only on materials from the point of view of genetic status and relationships. In real projects, it should be combined with other ontologies for describing bibliographic and archival resources. In the Roud dataset, GENO has been combined with the project-specific ontology ROUD-OEUVRES, so that the complete previous example is as follows:

```
@prefix geno: <https://w3id.org/geno#>.
@prefix fabio: <http://purl.org/spar/fabio/>.
@prefix roud-oeuvres: <https://api.ls-prod-server.dasch.
swiss/ontology/0112/roud-oeuvres/v2#>.

<https://ark.dasch.swiss/ark:/72163/1/0112/
QKyBsUJoRyW4JphDsWdRLQ8> a roud-oeuvres:Manuscript,
geno:Avant-textualWitness;
  roud-oeuvres:manuscriptIsInArchive "CLSR GR";
  roud-oeuvres:manuscriptHasShelfmark "MS 1 A/1b";
  roud-oeuvres:manuscriptHasTitle "Pardonne à ce cœur
dur, Bain";
  geno:hasGeneticStage geno:ClearCopy;
  geno:isMemberOfDossier

<https://ark.dasch.swiss/ark:/72163/1/0112/
ctZuEa2LTL=PmBoYPWABXQY>.
<https://ark.dasch.swiss/ark:/72163/1/0112/
ctZuEa2LTL=PmBoYPWABXQY> a geno:GeneticDossier;
  geno:resultsInPublication
<https://ark.dasch.swiss/ark:/72163/1/0112/RzaFM7
DjQDGal0kKZU7UCgt>.

<https://ark.dasch.swiss/ark:/72163/1/0112/
RzaFM7DjQDGal0kKZU7UCgt> a fabio:Expression,
roud-oeuvres:PeriodicalArticle, geno:Publication;
  roud-oeuvres:isPublishedInPeriodical
<https://ark.dasch.swiss/ark:/72163/1/0112/
QnH_XlxWT6uv28PmahXbYgw>;
  roud-oeuvres:publicationHasDate "1935-02";
  roud-oeuvres:publicationHasTitle "Bain d'un faucheur";
  geno:publicationIsReusedInDossier
<https://ark.dasch.swiss/ark:/72163/1/0112/
uFwo=VxATUy=mhxashFzbwf>.

<https://ark.dasch.swiss/ark:/72163/1/0112/QnH_
XlxWT6uv28PmahXbYgw> a roud-oeuvres:Periodical;
  roud-oeuvres:periodicalHasTitle "Présence".
```

4. Methods

This section comments on the methods of data collection and analysis.

The Roud dataset was created as part of the above-mentioned project "Gustave Roud, *Œuvres complètes*". The main output of the project was a printed critical edition of Roud's complete works (Roud, 2022), edited by Julien Burri, Alessio Christen, Claire Jaquier, Raphaëlle Lacord, Daniel Maggetti, Bruno Pellegrino and Stéphane Pétermann. Elena Spadini, the author of this article, was responsible for encoding the information in the print critical edition into the RDF dataset, manually and via scripts using the GUI and the API of the DaSCH Service Platform, as documented in the "Technical documentation" page in *Gustave Roud. Textes & Archives* (Jaquier & Maggetti, 2023). As mentioned above, the data presented here is a subset of the full project data: this subset concerns only the genetic networks and is extracted from the full dataset via SPARQL queries available at www.github.com/gustaveroudproject/geneticNetworksDataViz/tree/master/doc. The project data was modelled according to version 0.2 of the GENO ontology. After the end of the project, a new version of the ontology (1.0, www.w3id.org/geno) was developed and the dataset was ported to this version via scripting.

The much smaller Comment dataset was manually created based on the direct observation of the genetic witnesses and on the archival documentation available at the Swiss Literary Archives, where the original documents are kept.

Various SPARQL queries are provided together with the datasets at www.github.com/gen-o/geno/tree/master/current/dataSample. They can be run in a triplestore or using the Jupyter notebook provided with the documentation, which can also be executed in a virtual machine via the mybinder.org service to avoid local installations. The same queries can be easily adapted to specific datasets: in the example below, it is sufficient to add or remove the commented line. This SPARQL query returns all witnesses belonging to the pre-compositional phase and their genetic stages:

```
prefix geno: <https://w3id.org/geno#>
# prefix roud-oeuvres: <http://www.knora.org/
ontology/0112/roud-oeuvres#>

select * where {
    ?s a geno:PrecompositionalPhaseResource.
    # ?s roud-oeuvres:manuscriptHasTitle ?title.
    ?s geno:hasGeneticStage  ?stage.

}
```

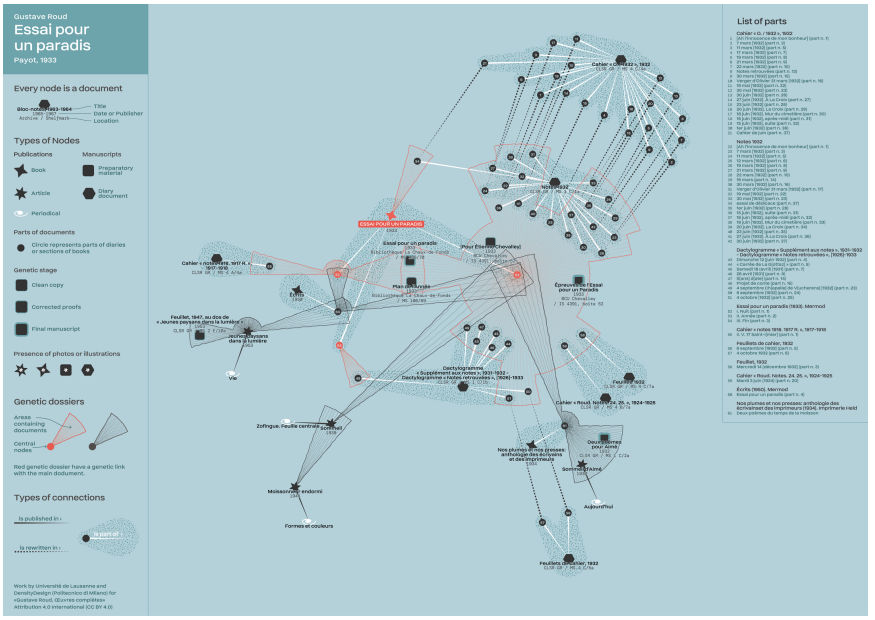


FIGURE 2 Data visualisation of the genetic network of "Essai pour un paradis" by Gustave Roud

The Roud dataset was used to create ten data visualisations in collaboration with the designers from the Density Design Lab at the Politecnico di Milano (Elli et al., 2023). The data for each book in JSON-LD was converted into a table of nodes and edges and rendered visually as a network, finalised using vector editing software. In contrast to the visualisation discussed above (see section 'Problem'), these are data visualisations, that visually and semi-automatically represent the structured data created by the scholars. The visualisations make it possible to graphically identify characteristics of Gustave Roud's genetic profile, such as the reuse from the diary and from published texts. Diary reuse, for example, is shown by the accumulation of small circles, while the rewriting of diary notes from one support to another is represented in the "marionette" structure (see Figure 2). The ten data visualisations are published in the web application *Gustave Roud. Textes & Archives* (Jaquier & Maggetti, 2023).

5. Concluding Remarks

The availability of structured data in open datasets in the field of genetic criticism and authorial philology is a novelty that could open up new

perspectives and allow new research questions to be asked. Pierre-Marc De Biasi argues that defining a "functional typology of genetic materials" is necessary because "it is by means of such a definition that each genetic profile can be described" (1996, p. 36). Allowing for genetic profiles to emerge is also one of the main goals of modelling the genetic information and of producing datasets such as those presented in this article: within the same author and between authors, profiles can provide a different entry point into the writing process of major and minor literary works.

In conclusion, the datasets presented in this article make explicit in the form of computable resources an important and often implicit part of the genetic analysis, the so-called *classement des manuscrits*. Such datasets can be created even when the texts cannot be made available due to copyright or other restrictions. The presentation of the datasets in this paper will hopefully encourage new reuse scenarios and the creation of other structured datasets in this research area.

References

Dängeli, P., Wieland, M., Wirtz, I. M., & Zumsteg, S. (Eds.). (2016). *Hermann Burger: Lokalbericht. Digitale Edition*. www.lokalbericht.ch.

De Biasi, P.-M. (1996). What is a literary draft? Toward a functional typology of genetic documentation. *Yale French Studies*, *89*, 26–58. www.doi.org/10.2307/2930337.

De Biasi, P.-M. (2008). Les six grandes étapes de la recherche en génétique des textes. In A. Crasson (Ed.), *L'édition du manuscrit: De l'archive de création au scriptorium électronique* (pp. 25–46). Editions Academia.

Dillen, W., and Neyt, V. (2016). Digital Scholarly Editing within the Boundaries of Copyright Restrictions. *Digital Scholarship in the Humanities*, *31*(4), 785–796. www.doi.org/10.1093/llc/fqw011.

Elli, T., Benedetti, A., Pallacci, V., Spadini, E., & Mauri, M. (2023). Designing network visualizations for genetic literary criticism. *Convergences – Journal of Research and Arts Education*, *16*(31), 25–38. www.doi.org/10.53681/c1514225187514391s.31.176.

Girard, D., Leclerc, Y., & Durel, M. (Eds.). (2009). *Les manuscrits de Madame Bovary. Édition intégrale sur le web*. www.bovary.fr.

Italia, P., & Raboni, G. (2021). *What is authorial philology?* Open Book Publishers. www.doi.org/10.11647/obp.0224.

Jaquier, C., & Maggetti, D. (Eds.). (2023). *Gustave Roud. Textes & Archives*. https://roud.unil.ch.

Johansson, F., & équipe Valery (ITEM) (Eds.). (2018). *Robinson de Paul Valéry: Édition génétique*. www.eman-archives.org/valery-robinson.

Probst, R., & Weber, U. (Eds.). (2021). *Friedrich Dürrenmatt. Das Stoffe-Projekt*. www.fd-stoffe-online.ch.

Roud, G. (2022). *Œuvres complètes* (C. Jaquier & D. Maggetti, Eds.). Zoé.

Spadini, E. (Ed.). (2023a). *Bernard Comment, "En mer". A scholarly edition prototype*. https://comment-enmer.github.io.

Spadini, E. (2023b). *GENO, the Genetic Networks Ontology* (1.0). www.w3id.org/geno.

Spadini, E., Christen, A., Pallacci, V., Elli, T., Benedetti, A., Maggetti, D., Mauri, M., & Pétermann, S. (2023). Genetic networks: Data model and visualisations. In A. Baillot, T. Tasovac, W. Scholger, & G. Vogeler (Eds.), *Digital Humanities 2023: Book of Abstracts*. Graz. www.doi.org/10.5281/zenodo.8210808.

Van Hulle, D. (2022). *Genetic criticism: Tracing creativity in literature* (1st ed.). Oxford University Press. www.doi.org/10.1093/oso/9780192846792.001.0001.

Van Hulle, D., & Nixon, M. (Eds.). (2011–2022). *Samuel Beckett Digital Manuscript Project*. University Press Antwerp (ASP/UPA). www.beckettarchive.org.