e-ISSN 2724-3923

magazén
Vol. 6 – Num. 2 – December 2025

The Genetic Dossier in the Web of Data
From Documentary Collections to a Scholarly Archive

Elsa Pereira
University of Porto, Faculty of Arts and Humanities, CITCEM, Portugal

Abstract While archivists and genetic scholars differ considerably in their methodological frameworks, the digital turn in archival preservation and scholarly editing provides an opportunity to narrow the gap. This article examines how Semantic Web technologies can bridge differing approaches to documentary collections of contemporary authors, while also outlining two current challenges to this pursuit: some limitations of LOD in representing genetic dossiers in informative ways and a series of legal issues that prevent digital scholarly archives of genetic orientation from realising their full potential in the Web of Data.

Keywords Archive. Recordkeeping. Scholarly editing. Genetic dossier. Semantic web. Copyright.

Summary

This article was supported by Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the scope of UIDB/04059/2025 (DOI: 10.54499/UID/04059/2025).

Edizioni
Ca'Foscari

Peer review

Submitted 2025-09-29
Accepted 2025-11-28
Published 2025-12-23

Open access

Citation Pereira, E. (2025). “The Genetic Dossier in the Web of Data”. *magazén*, 6(2), 1-16.

1 Introduction

Archivists and genetic scholars differ considerably in their methodological frameworks, which may explain why they

> are largely not taking part in the same conversations, not speaking the same conceptual languages, and not benefiting from each other's insights. (Caswell 2016, 2)

The digital turn in archival preservation and scholarly editing presents a window of opportunity to bring the two fields together. As more GLAM institutions responsible for preserving and providing access to literary archives of modern or contemporary authors undertake large-scale digitisation of material under their custody, new possibilities emerge for digital scholarly projects to reconnect those resources within a Semantic Web ecosystem, fostering closer collaborations between genetic scholars and archival institutions. However, the Web of Data also presents specific challenges that warrant further research and attention.

After comparing differing organising approaches to literary archives in recordkeeping and genetic criticism, this article will consider how Semantic Web technologies may bridge the methodological frameworks of both fields, while outlining two current obstacles to this pursuit: some limitations of LOD in representing genetic dossiers in informative ways and a series of legal issues that prevent digital scholarly archives of genetic orientation from realising their full potential in the Web of Data.

2 Literary Archives in Recordkeeping vs. Genetic Criticism

Typically comprising the working and personal papers of modern or contemporary authors, literary archives are characterised by a wide variety of material,[1] whose significance and appreciation set them apart from most other types of archival repositories. Increasingly cherished by literary enthusiasts and collectors since the beginning

[1] "The ideal literary collection captures the full gamut of a writer's work – background notes and research; annotated books and critical editions; literary drafts; photographic components; audio material; personal journals; literary logs; objects like keepsakes or awards; correspondence with publishers, editors and friends; editors' and printers' proofs, and final copies" (Molloy 2019, 328). The Group for Literary Archives & Manuscripts at the University of Manchester (GLAM) and the Group for Literary Archives and Manuscripts – North America (GLAMNA) have further systematised the material typically found in literary archives: `http://glam-archives.org.uk/?page_id=1731` (12/12/2022).

of the Romantic *Geniezeit*,[2] the relevance of literary archives lies in "the insights they give into the act of creation", which translates into "a higher financial value" and may explain why an author's holographs end up being "scattered in diverse locations" worldwide (Sutton 2014, 295-6):

> Literary archives […] tend to travel much further than other types of papers and to be housed in unpredictable locations – often […] determined by market forces rather than by internal archival logic. […L]iterary papers are usually found […] to be divided between several collecting institutions. This phenomenon, which we have come to call 'split collections' […] represent[s] an essential part of the world of literary manuscripts. (Sutton 2018, 7-8)

Indeed, very few writers have their manuscripts entirely preserved in a single institution, whether a public or university library, a state archive, a private foundation, a literary house, or a museum. Not only do authors themselves disseminate documentary evidence of their writing, offering drafts as a memento or gift of friendship (Boie 1993, 42-43) to those with whom they correspond throughout their lives, but, after passing away, their estates are also frequently divided among heirs and subject to the sort of posthumous plunder that is implicit in the Latin root of the Portuguese term designating literary archives: *espólio* (from the Latin *spolia*, 'spoils' or 'stolen treasures').

While archival studies and literary genetic criticism approach the *spoils* of an author with very different methodologies, we believe their perspectives and understandings "overlap and can be brought together" (Bunn, Rayner 2019, 360).

Archival studies is a subfield of information science dedicated to "the nature, management, and uses of records", defined as "persistent representations of activities, created by participants or observers" (Caswell 2016, 3, 5). Traditionally, recordkeepers curating literary archives are guided in their work by a foundational principle of "respect for the fonds" (Muller, Feith, Fruin 2003, 54), which implies preserving the so-called "provenance" of archival documents that bear an organic relationship to one another. In practice, this means that documentary pieces shall not be aggregated by subject or any

[2] With the privatisation of intellectual property and the concomitant valorisation of the creative genius during the Romantic period, writers started to preserve draft manuscripts systematically. However, it was not until the second half of the twentieth century that extensive literary archives proliferated, as "increasing attention was being given to the processes of literary composition and revision in their own right. At the forefront of this practice were French scholars associated with the Centre d'Analyse des Manuscrits in Paris, formed after the accession of Heinrich Heine's papers by the Bibliothèque Nationale in 1966" (Anderson et al. 2021, 8).

other interpretative criteria; instead, records created by different individuals must be "kept separately", with their "original order" and context preserved (Caswell 2016, 7), while classified according to genre, type, or material support (Lopes 2007, 55).

This traditional understanding of "provenance" as an organising principle of archival studies contrasts with the "speculative approach" (Drucker, Nowviskie 2004, 431) of scholarly editors with a "genetic orientation" (Van Hulle, Shillingsburg 2015, 36), who, conversely, view records not as a "fixed product" (Bunn, Rayner 2019, 369), but as "dynamic objects in motion" (Caswell 2016, 6).

Originating in France during the 1960s, in connection with the Centre d'Analyse des Manuscrits in Paris (later evolved in the current ITEM – Institut des Textes et Manuscrits Modernes), genetic criticism succeeded in adding a temporal, paradigmatic dimension to the literary text, regarded as a process rather than a product, by drawing attention to its variations in draft form and all the transformations that result from the author's writing or rewriting activity over time. This type of "archaeology of the manuscript" (Van Mierlo 2013) relies on literary archives, mainly from the nineteenth century onward, to provide insight into the compositional development of a literary work and expand the interpretation possibilities of the text:

> manuscripts […] offer up new and unseen material, and also suggest, in their very physicality, the writing methods and processes unique to the subject of study. They can further 'solve factual problems like the dating of a poem or establishing an accurate text' and 'illuminate the broader meanings of a literary work' (Gioia 2004: 36). Beyond this, archival materials offer us other conduits of research and knowledge, […revealing], as Cook argues, the 'context behind the text, the power relationships shaping the documentary heritage, and indeed the document's form and content'. (Stead 2016, 4)

For that to be in place, scholars must compile a *genetic dossier* (Grésillon [1994] 2016, 286) comprising all the physically dispersed documents of an author's writing project that "bear witness to the evolution of the work" (De Biasi 2004, 38). This may include the version records that preceded publication (e.g. notes, drafts, revised manuscripts, typescripts, print proofs), as well as other correlated evidence of the broader interpersonal networks that contribute to the author's creative process, such as his library and correspondence. The methodology involves not only collecting all extant genetic documentation (*recensio*) but also comparing the respective textual variants to infer the genealogical relationships among the collected pieces and organising the work's *avant-texte* (Bellemin-Noël 1972) according to the writing chronology.

Whereas recordkeepers use systematic guidelines in their archival practice, based on such principles as provenance, collective control, or original order of records, genetic scholars aggregate different authorial material of various provenance (e.g. marginalia in books, notebooks, draft manuscripts, typescripts, letters exchanged with other people), subjectively organising the jigsaw pieces into speculative archives – the *genetic dossier* – aimed at reconstructing the author's creative process.

In recent years, the digital medium has significantly facilitated the constitution of these interpretative *scholarly archives*,[3] allowing researchers to aggregate facsimiles and transcriptions of material scattered across different institutions and model their textual relations within a dedicated virtual environment for specific academic purposes.

3 Transitioning Genetic Dossiers into the Web of Data

Over the past decade, scholarly editorial initiatives, such as the Samuel Beckett Digital Manuscript Project,[4] the Shelley-Godwin Archive,[5] or the Gustave Roud: Textes & Archives,[6] have succeeded in digitally reuniting dispersed documentation of modern and contemporary authors, facilitating the examination of the genetic dossier of their works.

Although Wout Dillen has rightly noted that many of these

[3] For more systematic definitions of "scholarly archive" in DH projects, distinguishing it from the traditional notion of "archive" in archival studies, see e.g. Theimer 2012; Adema, Stoyanova 2015.

[4] Van Hulle, Nixon 2011-present. The project was developed by the Centre for Manuscript Genetics at the University of Antwerp, the Beckett International Foundation at the University of Reading, the Oxford Centre for Textual Editing and Theory at the University of Oxford, and the Harry Ransom Humanities Research Center at the University of Texas at Austin, with the permission of the Estate of Samuel Beckett.

[5] Fraistat et al. 2013-present. The project aims to unite online the widely dispersed handwritten legacy of Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft. It is the result of a partnership between the New York Public Library and the Maryland Institute for Technology in the Humanities, in cooperation with Oxford's Bodleian Library, the Huntington Library, the British Library, the Houghton Library, and the Victoria and Albert Museum.

[6] Jaquier, Maggetti 2022. Developed at the University of Lausanne and supported by the Swiss National Science Foundation (2017-2022), the project provides a critical print edition of Gustave Roud's complete works, and a genetic digital archive of the authorial material housed at the Centre des Littératures en Suisse Romande..

DH projects call themselves *archives*,[7] their "archival impulse" (Eggert 2019) diverges from the principles followed by librarians and recordkeepers, who are invariably bemused by the term used in this context.[8] Instead of attending to the provenance of documents, digital scholarly archives of genetic orientation are "hermeneutical instruments" (Ramsay, Rockwell 2012, 79) aimed at organising a "purposeful collection of surrogates" (Price 2008) to reveal interpretative connections between dispersed textual witnesses. Presenting themselves as a "work-site" (Eggert 2005, 433), a "knowledge site" (Shillingsburg 2006, 88), or a "platform for learning" (Theimer 2014, 146), those projects actively engage with the dynamics of variation through a range of digital tools to allow readers to navigate multiple versions of a text and follow the author's compositional development over a more or less extended period of experimentation and revision.

In addition to XML-TEI markup and algorithmic collation, participatory editorial projects, such as the LdoD Archive,[9] have been exploring social editing functionalities, supported by structured databases and Web 2.0 environments that enable users to create *virtual editions* of the documentation.[10] Scholars are invited to manipulate the "dynamic layer of the archive" (Portela 2022, 191), either performing editorial script acts such as annotations, or reconfiguring the writing sequence of texts, in what some have also been labelling as new interactive "forms of analysis and creativity", in line with the poststucturalist "esthetic of the possible" (Gooding et al. 2019, 386, 376).

[7] "[…] projects that are generally considered as digital scholarly editions often do not shy away from calling themselves archives […] – think, for instance, of the *William Blake Archive*, the *Piers Plowman Electronic Archive*, the *Walt Whitman Archive*, and, more recently, the *Shelley-Godwin Archive*. […] As the digital medium started to break down the borders between archives and editions, […] the user can decide how to use the digital resource: as an archive of textual documents and image reproductions; as a (genetic) dossier that organises these documents and exposes their internal logic; or as an edition, a curated and edited collection of texts that informs the reader on the textual tradition of the work" (Dillen 2019, 265, 267).

[8] As Kate Theimer observed, "[a]rchivists would not refer to online groupings of digital copies of non-digital original materials, often comprised of materials (including published materials) located in different physical repositories or collections, purposefully selected and arranged in order to support a scholarly goal, as an 'archives' – and so the confusion of an Archivist tourist in the land of Digital Humanities" (Theimer 2012).

[9] Portela, Silva 2017-present. Developed at the Centre for Portuguese Literature at the University of Coimbra and funded by the Portuguese Foundation for Science and Technology and the European Regional Development Fund, the LdoD Archive is a collaborative digital archive of the *Book of Disquiet* by Fernando Pessoa. It contains images of the autograph documents, transcriptions of those documents, and also transcriptions of four editions of Pessoa's work.

[10] See Silva, Portela 2015.

More recently, digital scholarly archives of genetic orientation have also been drawing inspiration from advancements in Linked Open Data (LOD) and other technologies that follow up on Berners-Lee's vision of a Semantic Web of Data (Berners-Lee et al. 2001; Berners-Lee 2006), to reveal and enhance the complex network of relationships devised among various documents of a genetic dossier. In a nutshell, these projects use persistent URIs to identify resources and apply web ontologies to formally represent relationships or the underlying logic among different nodes in the documentary network. Resource Description Framework (RDF) datasets, represented as subject-predicate-object triples, will model a graph structure that computers can interpret, while users can interact via a SPARQL endpoint with a graphical interface that leverages the semantic layer for querying and manipulating the graph database.[11]

Among the projects that have been applying Semantic Web technologies to genetic dossiers,[12] the *Gustave Roud: Textes & Archives* (Jaquier, Maggetti 2022) deserves special mention, as the team designed a new data model, formalised in the Web Ontology Language, specifically for Genetic Criticism.[13] This GeNO ontology effectively describes the interwoven networks within and outside an author's genetic dossier and can be queried using cURL and Gravsearch, a virtual graph search based on SPARQL that allows researchers to find, for instance, which diary entry of the author ended up in his fictional work.

The *Shelley-Godwin Archive* (Fraistat et al. 2013-present) is another initiative worth mentioning, as it builds on linked data principles and the Shared Canvas data model to support a participatory platform where anyone on the web can describe, discuss, and reuse facsimiles and transcriptions of archival material, within a global, interconnected network of information that aligns with the 5S model[14] and an interdisciplinary vision of "Linked Research" (Capadisli 2016).

[11] For a comprehensive perspective on graph data-models and Semantic Web technologies in scholarly digital editing, see Spadini, Tomasi, Vogeler 2021.

[12] A noteworthy project developed in Italy is the digital edition of Paolo Bufalini's notebook (Daquino et al. 2020).

[13] Geno – the Genetic Networks Ontology (Spadini 2023), which builds as an extension of the knora-base ontology. Other existing ontologies for semantic editions, such as CAO – Critical Apparatus Ontology (Giovannetti 2019) and CEO – Critical Edition Ontology (Martignano 2023) did not adequately describe the network of textual witnesses in relation with other witnesses in the genetic dossier. See Christen, Spadini 2019, 84.

[14] The 5S model refers to the fundamental concepts of Streams, Structures, Spaces, Scenarios, and Societies (5S) that formally model digital libraries, regarded as "a managed collection of information with associated services involving communities where information is stored in digital formats and accessible over a network" (Gonçalves 2004, 19). For a comparative approach between the frameworks of digital libraries, archives, and editions, see e.g. Meschini 2020, chapter 3.

Interestingly, the project stems from a partnership with several public and university libraries, expanding still-rare collaborations between literary scholars and recordkeepers[15] into the Web of Data and opening up possibilities for further cooperation.

In fact, as more GLAM institutions lead large-scale digitisation projects that adhere to protocols such as the International Image Interoperability Framework (IIIF) and the Text Encoding Initiative (TEI), new opportunities emerge for archivists to incorporate digital scholarly archives into local descriptions, enhancing or contextualising their records. Conversely, scholars should also be able to connect genetic dossiers to the authors' archival repositories and personal libraries,[16] allowing users to navigate the virtual research interface without losing contact with the provenance and archival order of the material records. But while steps have been taken along the path, the vision of a global web of literary archives remains "far away on the utopian horizon" (Fordham, cited in Anderson et al. 2021, 7), hindered so far by at least two main obstacles.

The first issue that stands out is the lack of shared vocabularies, ontologies, and "good human-usable interfaces for the Semantic Web" (Brown, Simpson 2015). Recent initiatives promoting Linked Open Data vocabularies for the description of manuscripts[17] and textual

[15] "Think for example of *Litteraturbanken*, the 'Swedish Literature Bank' […]. In an impressive collaborative effort between literary and linguistic scholars, research libraries, and editorial societies and academies, this project contains a wide range of digital facsimiles and their (corrected OCR based) transcriptions of documents pertaining to Swedish literary works from the Middle Ages to the present. Alongside their edited texts available in HTML (and, when possible, EPUB), these are contextualised further by means of scholarly introductions, presentations, other didactic materials, and even allow for basic text analysis functionalities through a collaboration with Språkbanken, the 'Swedish Language Bank'" (Dillen 2019, 265).

[16] Many twentieth-century authors have not only their manuscripts but also their libraries preserved and digitised. Among other examples, it is worth mentioning the Private Library of Portuguese author Fernando Pessoa (1888-1935), comprising roughly 1300 books once owned by the poet and currently available on the website of Casa Fernando Pessoa: `https://bibliotecaparticular.casafernandopessoa.pt/index/index.htm`. Although the fully digitised collection has been available for several research projects on Pessoa's marginalia, the digital library is not IIIF-compliant, which makes it not ideal to incorporate within a semantic web environment.

[17] See e.g. the efforts developed by the Working Group for Linked Manuscript Descriptions, whose goal is to create a common Linked Open Data vocabulary for the description of medieval manuscripts from the Middle East. The working group met for its first sessions online on 15-16 December 2021 as part of the Linked Pasts VII Symposium, hosted by Ghent University. `https://www.ghentcdh.ugent.be/linked-pasts-vii-symposium`.

variation[18] are promising contributions to addressing the problem, but so far, digital scholarly projects experimenting with ontologies to express textual relationships across different authorial materials have designed their data models independently of the archival records underpinning the projects.[19] One suggestion to overcome the current disconnect between literary archives and genetic dossiers, enabling rich, interlinked data to be shared and repurposed by third-party applications, could involve open knowledge bases such as Wikidata, as well as Solid Pods specifically designed for GLAM institutions to share archival records and promote decentralised data networks, as this kind of resource enables different web APIs to provide new views into the knowledge graph.[20] Still, while knowledge graph visualisation for the Solid ecosystem is making progress and paving the way for further research,[21] experts in graph technologies recognise that network graphs in general do not present complex datasets of textual information in a clear and intelligible manner, mainly because Linked Data is a machine-readable format not intended for humans,[22] and "[m]any levels of discursive mediation are needed for the methods of close and distant reading to productively inform one another" (Stoyanova 2023, 39). The *Gustave Roud: Textes & Archives*, for instance, drew inspiration from celestial maps to explore the centrality of the author's diary within the genetic dossier, achieving positive usage test results among experts (Elli et al. 2023, 32, 36), but interpreting graph representations of such complex data

[18] In this regard, see Bleeker et al. 2025. The authors have also recently established a working group on Visualizing and Investigating Differences In Texts (VIDIT), aimed at building a global community of scholars, developers, and designers interested in studying and visualising variation in historical and literary texts. `https://wg-vidit.github.io/`.

[19] The *Gustave Roud: Textes & Archives'* data model, for instance, is independent of the archival online inventory, available at the Centre for Literatures in French-speaking Switzerland: `https://atom-archives.unil.ch/index.php/ch-000225-8-p73`.

[20] Solid (SOcial LInked Data) is a set of technology specifications for the Web of Data, which includes decentralised online pods, ("often referred to as data vaults), standard communication between apps, and the use of a universal data format in the form of a Resource Description Framework (RDF) […]. The central notion of Solid is the technical and organizational separation of data, services, and identity. In this way, Solid as a set of technology specifications enables the creation of decentralized applications using W3C standards and protocols […], which counterweights the current dominant architecture of the internet" (Theys et al. 2025, 505).

[21] In this regard, see e.g. Dedecker et al. 2022.

[22] "ancora tanta strada c'è da fare nella realizzazione di applicazioni che siano in grado di utilizzare in modo sapiente quei dati per restituire all'utente sotto forma di nuova conoscenza. Partiamo dal presupposto che i LOD non sono pensati originariamente per l'utente, ma per l'elaborazione da parte della macchina" (There is still a long way to go in the realisation of applications to wisely use data and return it to the user in the form of new knowledge. LOD is not originally designed for the user but for processing by the machine) (Tomasi 2022, 132-3).

sets is incredibly difficult for literary scholars without the technical knowledge for querying the ontology.[23] As such, we need archivists, genetic scholars, and data scientists to come together and codevelop graphic user interfaces that make graph network visualisation more accessible and “informative for a wider audience”.[24]

In the case of modern and contemporary literary archives, however, another major obstacle to reconnecting authorial material within a Semantic Web ecosystem stands out, due to a fundamental conflict between the free availability of distributed data, implicit in the concept of LOD,[25] and a series of legal restrictions affecting authorial repositories, particularly in countries with a legal tradition of *Droit d’Auteur*. Despite usually being deposited in institutions funded by public resources, manuscripts and other archival material of twentieth and twenty-first-century writers is subject to a series of copyright and non-copyright restrictions that protect the privacy and moral rights of authors, forcing GLAM institutions to “curtail the widespread digitisation of whole collections” (Anderson et al. 2021, 7) and restrict access to “a minority of researchers who have the time and funding” to view the documents on site (Jaillant 2019, 290). In fact, those willing to use contemporary literary materials for research purposes often find themselves in a never-ending maze of bureaucracy, including formal authorisations from both copyright owners and custodians of the material, which implies dealing with different propriety, authority, dependency, and privacy restrictions:

> The owner of copyright for material in the Manuscripts Collection is the writer or creator of the material, or the creator’s legal heir(s). Note that the donor of the material is not always the copyright owner. In addition, many collections contain a variety of letters, diaries, documents owned by multiple copyright owners. […] Should you wish to publish material from the Library’s Manuscript Collection, you will need to: declare your intention to the Library

[23] See e.g. “Constellation génétique de Campagne perdue de Gustave Roud”. `https://roud.unil.ch/resources/http%3A%2F%2Frdfh.ch%2F0112%2FpKBOXI-GSEyVBBECW1Xgkw`. A simple way to improve the legibility of this genetic network would be incorporating weblinks to the different nodes connected in the graph, allowing users to navigate the digital scholarly archive taking the celestial map as a reference.

[24] Statement issued by the VIDIT: `https://wg-vidit.github.io/`. For an overview of current graph visualisation techniques, specifically applied to collation outputs, see Birnbaum, Dekker 2024.

[25] Linked Open Data is a combination of two basic concepts: linked data (a method of storing information based on the connections and relationships between items) and open data (signifying data that has been made freely available for distribution).

> as custodian of the material, obtain copyright clearance from the copyright holder(s).[26]

As demonstrated in a previous article dedicated to major legal obstacles encountered by European genetic scholars, recent exceptions introduced by the CDSM Directive did not meet the requirements of ongoing advancements in digital humanities,[27] making the path towards publishing and providing online international access to contemporary literary archives difficult to navigate, especially for unpublished works, where “the waters are particularly muddy” (Dillen, Neyt 2016, 788). Before taking further steps towards a much-anticipated vision of genetic dossiers in the Web of Data, scholars investigating twentieth and twenty-first-century authors therefore need bold policy-making adjustments to ensure that their work will not be rendered worthless by someone refusing publication permission:

> we need to extend the scope of the available exceptions […] to allow for scholarly publication in the digital age – or otherwise, a legal license designed with scholarship in mind so that academic researchers may work with published texts and holographic materials in public archive libraries, disclosing research results (in person, on paper, and online) without interference from heirs or successors. Moreover, we also need national or European management systems led by independent copyright boards to facilitate the clearance of orphan works for different uses and reduce the randomness of our current authorisation system. (Pereira 2023, 523-4).

[26] National Library of Australia, “Rights and the Manuscripts Collection”. https://www.library.gov.au/services/copyright-library-collections/rights-and-manuscripts-collection.

[27] European Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market. Digital literary approaches affected by the TDM exception include: “classification and clustering of texts (e.g. for authorship attribution and stylometry), extraction of distinctive features, semantic analysis with topic modelling, analysis of polarity with sentiment analysis, character relationships with network analysis, and analysis of relationships between texts (e.g. in text reuse). However, we should note that only those materials to which scholars have lawful access can be mined, and experiences in countries where TDM exceptions have been in force show that copyright issues will subsist: ‘Despite the TDM exception in German copyright law, Text and Data Mining (TDM) with copyrighted texts is still subject to restrictions, including those concerning the storage, publication and follow-up use of the resulting corpora’” (Pereira 2023, 521-2).

4 Conclusion and Outlook

In his 2014 book on memory and scholarship in the age of digital reproduction, Jerome McGann argued that to study literary creativity, scholars needed "cultural records to be comprehensive, stable, and accessible" while being able to augment "that record with our own contributions" (McGann 2014, 131-2). The proposal implied shifting the idea of archival records as fixed informational resources to embrace the digital scholarly perspective on the term, regarded as "a complex system inhabited by all the different agents involved in the production of academic work" (Bunn, Rayner 2019, 369).

In this regard, Semantic Web technologies present a window of opportunity to enhance interdisciplinary collaboration among archivists, textual scholars, and genetic critics, reapproaching their methodological frameworks to think anew about the working methods of prominent writers. While the persistence of practical obstacles to this pursuit leaves the full potential of genetic dossiers in the Web of Data untapped, much work has been done to overcome those shortcomings and interdisciplinary working groups must be formed to carry on the efforts, co-designing archival software for the Semantic Web ecosystem and simultaneously allowing for so-called "digital forensic work" on literary archives, i.e.:

> multiple modes of ordering and interpreting while also, at the same time, securing the collections that underpin this innovative work. (Gooding et al. 2019, 376)

Meanwhile, scholars and recordkeepers must come together and exert pressure on legislators to introduce the policy changes necessary to allow greater freedom in using copyrighted works for the preservation of cultural heritage in the Digital Single Market. The recent (albeit insufficient) TDM exception introduced into European legislation demonstrated that only through sustained commitment can we achieve the legal measures necessary to enable ongoing developments in computational literary studies. Allowing contemporary documentary collections and genetic dossiers to transition into the Web of Data should be our next goal.

Bibliography

Adema, J.; Stoyanova, S. (2015). “The Multidimensional Scholarly Archive”. *Open Reflections*. https://openreflections.wordpress.com/2015/10/28/the-multidimensional-scholarly-archive/.

Anderson, L.; Byers, M.; Warner, A. (2021). “Introduction: Poetry, Theory, Archives”. Anderson, L.; Byers, M.; Warner, A. (eds), *The Contemporary Poetry Archive: Essays and Interventions*. Edinburgh: Edinburgh University Press, 1-24.

Bellemin-Noël, J. (1972). *Le Texte Et l’Avant-Texte: Les Brouillons d’Un Poème De Milosz*. Paris: Larousse.

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001). “The Semantic Web: A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities”. *Scientific American*, 284, 1-5. https://www.researchgate.net/publication/225070375_The_Semantic_Web_A_New_Form_of_Web_Content_That_is_Meaningful_to_Computers_Will_Unleash_a_Revolution_of_New_Possibilities.

Berners-Lee, T. (2006). “Linked Data”. www.w3.org/DesignIssues/LinkedData.html.

Birnbaum, D.J.; Dekker, R.H. (2024). “Visualizing Textual Collation: Exploring Structured Representations of Textual Alignment”. *Proceedings of Balisage: The Markup Conference 2024*. Balisage Series on Markup Technologies, vol. 29. https://doi.org/10.4242/BalisageVol29.Birnbaum01.

Bleeker, E.; Spadini, E.; Nava, B.; Oostveen, B.; Dekker, R.H. (2025). “‘Here’s strangeness.’ A Collaborative Approach to Visualising Textual Variation”. https://doi.org/10.5281/ZENODO.15387538.

Boie, B. (1993). “L’écrivain et ses manuscrits”. Hay, L. (ed.), *Les Manuscrits des Écrivains*. Paris: Hachette, 34-53.

Brown, S.; Simpson, J. (2015). “An Entity by Any Other Name: Linked Open Data as a Basis for a Decentred, Dynamic Scholarly Publishing Ecology”. *Scholarly and Research Communication*, 6(2). http://src-online.ca/index.php/src/article/view/212/409.

Bunn, J.; Rayner, S.J. (2019). “Observing the Author-editor Relationship: Recordkeeping and Literary Scholarship in Dialogue”. *Archives and Manuscripts*, 47-3, 359-73. https://doi.org/10.1080/01576895.2019.1609363.

Capadisli, S. (2016). “Where is Web Science? From 404 to 200”. https://csarven.ca/web-science-from-404-to-200.

Caswell, M. (2016). “‘The Archive’ Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies”. *Reconstruction: Studies in Contemporary Culture*, 16(1). https://escholarship.org/uc/item/7bn4v1fk.

Christen, A.; Spadini, E. (2019) “Modeling Genetic Networks: Gustave Roud’s Oeuvre, from Diary to Poetry Collections”. *Umanistica Digitale*, 7, 77-104. https://doi.org/10.53681/c1514225187514391s.31.176.

Daquino, M.; Dello Buono, M.; Giovannetti, F.; Tomasi, F. (2020). *Paolo Bufalini: Appunti*. https://projects.dharc.unibo.it/bufalini-notebook/introduction.

De Biasi, P-M. (2004). “Toward a Science of Literature: Manuscript Analysis”. Deppman, J.; Ferrer, D.; Groden, M. (eds), *Genetic Criticism: Texts and Avant-textes*. Philadelphia, PA: University of Pennsylvania Press, 36-68.

Dedecker, R.; Slabbinck, W.; Wright, J.; Hochstenbach, P.; Colpaert, P.; Verborgh, R. (2022). “What’s in a Pod? A Knowledge Graph Interpretation for the Solid Ecosystem”. Saleem, M.; Ngonga Ngomo, A.-C. (eds), *Proceedings of the 6th*

Workshop on Storing, Querying and Benchmarking Knowledge Graphs, 81-96. https://solidlabresearch.github.io/WhatsInAPod/.

Dillen, W. (2019). "On Edited Archives and Archived Editions". *International Journal of Digital Humanities*, 1, 263-77. https://doi.org/10.1007/s42803-019-00018-4.

Dillen, W.; Neyt, V. (2016). "Digital Scholarly Editing Within the Boundaries of Copyright Restrictions". *Digital Scholarship in the Humanities*, 31-4. https://doi.org/10.1093/llc/fqw011.

Drucker, J.; Nowviskie, B. (2004). "Speculative Computing: Aesthetic Provocations in Humanities Computing". Schreibman, S.; Siemens, R.; Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell Publishing Professional, 431-47. https://doi.org/10.1002/9780470999875.ch29.

Eggert, P. (2005). "Text-Encoding, Theories of the Text, and the 'Work-Site'". *Literary & Linguistic Computing*, 20(4). http://doi.org/10.1093/llc/fqi050.

Eggert, P. (2019). "The Archival Impulse and the Editorial Impulse". *Variants*, 14, 3-22. http://doi.org/10.4000/variants.570.

Elli, T.; Benedetti, A.; Pallacci, V.; Spadini, E.; Mauri, M. (2023). "Designing Network Visualizations for Genetic Literary Criticism". *Convergências*, 16(31). https://doi.org/10.53681/c1514225187514391s.31.176.

Folsom, E. (2007). "Database as Genre: The Epic Transformation of Archives". *PMLA*, 122(5), 1571-9. https://doi.org/10.1632/pmla.2007.122.5.1571.

Fraistat, N.; Viglianti, R.; Denlinger, E.C.; (dir.) (2013-present). *The Shelley-Godwin Archive*. http://shelleygodwinarchive.org.

Gooding, P.; Smith, J.; Mann, J. (2019). "The Forensic Imagination: Interdisciplinary Approaches to Tracing Creativity in Writers' Born-digital Archives". *Archives and Manuscripts*, 47-3, 374-90. https://doi.org/10.1080/01576895.2019.1608837.

Giovannetti, F. (2019). *The Critical Apparatus Ontology (CAO)*. Version: 0.9. https://w3id.org/cao.

Gonçalves, M.A. (2004). *Streams, Structures, Spaces, Scenarios, and Societies (5S): A Formal Digital Library Framework and Its Applications*. Blacksburg: Faculty of the Virginia Polytechnic Institute and State University. https://www.academia.edu/54086046/Streams_structures_spaces_scenarios_and_societies_5S_A_formal_digital_library_framework_and_its_applications.

Grésillon, A. [1994] (2016). *Éléments de Critique Génétique*. Paris: CNRS Éditions.

Jaillant, L. (2019). "After the Digital Revolution: Working with Emails and Born-Digital Records in Literary and Publishers' Archives". *Archives and Manuscripts*, 47-3, 285-304. https://doi.org/10.1080/01576895.2019.1640555.

Jaquier, C.; Maggetti, D. (dir.) (2022). *Gustave Roud. Textes & Archives*. https://roud.unil.ch.

Lopes, F. (2007). "Como se trabalha no Arquivo de Cultura Portuguesa Contemporânea". *As Mãos da Escrita: 25 anos do Arquivo de Cultura Portuguesa Contemporânea*. Lisboa: Biblioteca Nacional de Portugal, 51-74. https://purl.pt/13858/1/abertura/como-trabalha-acpc.html.

Martignano, C. (2023). *Critical Edition Ontology (CEO)*. Version: 1.0. http://purl.org/critical-edition-ontology.

McGann, J. (2014). *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge: Harvard University Press.

Meschini, F. (2020). *Oltre il Libro: Forme di Testualità e Digital Humanities*. Milano: Editrice Bibliografica.

Molloy, K. (2019). “Literary Archives in the Digital Age: Issues and Encounters with Australian Writers”. *Archives and Manuscripts*, 47-3, 327-42. `https://doi.org/10.1080/01576895.2019.1631863`.

Muller, S; Feith, J.A.; Fruin, R. (2003). *Manual for the Arrangement and Description of Archives*. Transl. by A.H. Leavitt. 2nd ed. Chicago: Society of American Archivists.

Pereira, E. (2023). “Authors’ Rights vs. Textual Scholarship: A Portuguese Overview”. *Journal of Intellectual Property, Information Technology and E-Commerce Law*, 14-4, 510-24. `https://www.jipitec.eu/jipitec/article/view/19`.

Portela, M.; Silva, A.R. (dir.) (2017-present). *LdoD Archive: Collaborative Digital Archive of the Book of Disquiet*. `https://ldod.uc.pt/`.

Portela, M. (2022). *Literary Simulation and the Digital Humanities: Reading, Editing, Writing*. New York: Bloomsbury.

Price, K.M. (2008). “Electronic Scholarly Editions”. Siemens, R.; Schriebman, S. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell. `https://companions.digitalhumanities.org/DLS/?chapter=content/9781405148641_chapter_24.html`.

Ramsay, S.; Rockwell, G. (2012). “Developing Things: Notes toward an Epistemology of Building in the Digital Humanities”. Gold, M.K. (ed.), *Debates in the Digital Humanities*. Minneapolis: Minnesota Scholarship Online, 75-84. `https://doi.org/10.5749/minnesota/9780816677948.003.0010`.

Shillingsburg, P. (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.

Silva, A.R.; Portela, M. (2015). “TEI4LdoD: Textual Encoding and Social Editing in Web 2.0 Environments”. *Journal of the Text Encoding Initiative*, 8. `https://doi.org/10.4000/jtei.1171`.

Spadini, E. (2023). *GENO, the Genetic Networks Ontology*. Version: 1.0. `https://w3id.org/geno`.

Spadini, E.; Tomasi, F.; Vogeler, G. (eds) (2021). *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*. Norderstedt: Books on Demand.

Stead, L. (2016). “Introduction”. Smith, C.; Stead, L. (eds), *The Boundaries of the Literary Archive: Reclamation and Representation*. London; New York: Routledge, 1-13.

Stoyanova, S. (2023). “Articulating Intra- and Intertextual Relationships in the Fragment Collection. Working with the Digital Edition of Giacomo Leopardi’s Zibaldone”. *magazén*, 4(1), 13-42. `https://doi.org/10.30687/mag/2724-3923/2023/01/001`.

Sutton, D.C. (2014). “The Destinies of Literary Manuscripts, Past, Present and Future”. *Archives and Manuscripts*, 42(3), 295-300. `https://doi.org/10.1080/01576895.2014.948559`.

Sutton, D.C. (2018). “Introduction: Literary Papers as the most ‘Diasporic’ of all Archives”. Sutton, D.C., Livingstone, A. (eds), *The Future of Literary Archives: Diasporic and Dispersed Collections at Risk*. ARC Humanities Press.

Theimer, K. (2012). “Archives in Context and as Context”. *Journal of Digital Humanities*, 1(2). `https://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/`.

Theimer, K. (2014). “The Role of Archives in a Digital Society: Now is What Matters”. *Archivaria: The Journal of the Association of Canadian Archivists*, 78, 145-7. `https://archivaria.ca/index.php/archivaria/article/view/13498`.

Theys, T.; Mechant, P.; Maes, M.; Bourgeus, A.; Saldien, J.; De Marez, L. (2025). “Solid Pods: A Promising Approach to Enhance Users’ Perception of Data Transparency and Control”. *Interacting with Computers*, 37-6, 504-17. `https://doi.org/10.1093/iwc/iwaf017`.

Tomasi, F. (2022). *Organizzare la Conoscenza: Digital Humanities e Web semantico*. Milano: Editrice Bibliografica.

Van Hulle, D.; Nixon, M. (2021) "Editorial Principles and Practice". *Samuel Beckett Digital Manuscript Project*. https://www.beckettarchive.org/editorial.

Van Hulle, D.; Nixon, M. (dir.) (2011-present). *Samuel Beckett Digital Manuscript Project*. https://www.beckettarchive.org.

Van Hulle, D.; Shillingsburg, P. (2015). "Orientations to Text, Revisited". *Studies in Bibliography*, 37, 27-44. https://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv059.xml;chunk.id=d25715e2576;toc.depth=1;toc.id=d25715e2576;brand=default.

Van Mierlo, W. (2016). "The Archaeology of the Manuscript: Towards Modern Palaeography". Smith, C.; Stead, L. (eds), *The Boundaries of the Literary Archive: Reclamation and Representation*. London; New York: Routledge, 15-29.