MONOGRAPH
METODE SCIENCE STUDIES JOURNAL (2023). University of Valencia.
ISSN: 2174-3487. eISSN: 2174-9221.
Submitted: 10/11/2023. Approved: 29/01/2024.
https://doi.org/10.7203/metode.15.27672

WHERE WILL THE DIGITAL HUMANITIES BE IN 100 YEARS?

The humanities as a hope for the digital

José Calvo Tello

The acceleration of technological development raises certain challenges and doubts about the maintenance of tools and data. Must we accept that digital data will be useless in a few decades? This article presents some concrete solutions, such as the use of repositories, the application of FAIR principles, or the use of standard formats such as XML-TEI. More generally, it argues for the maintenance of a historical view both of the humanities and of libraries, and for a humanistic and quantitative critique of digital aspects.

Keywords: **digital humanities**, **data**, **preservation**, **FAIR principles**, **repositories**.

DIGITAL HUMANITIES, A NOT-SO-NEW FIELD

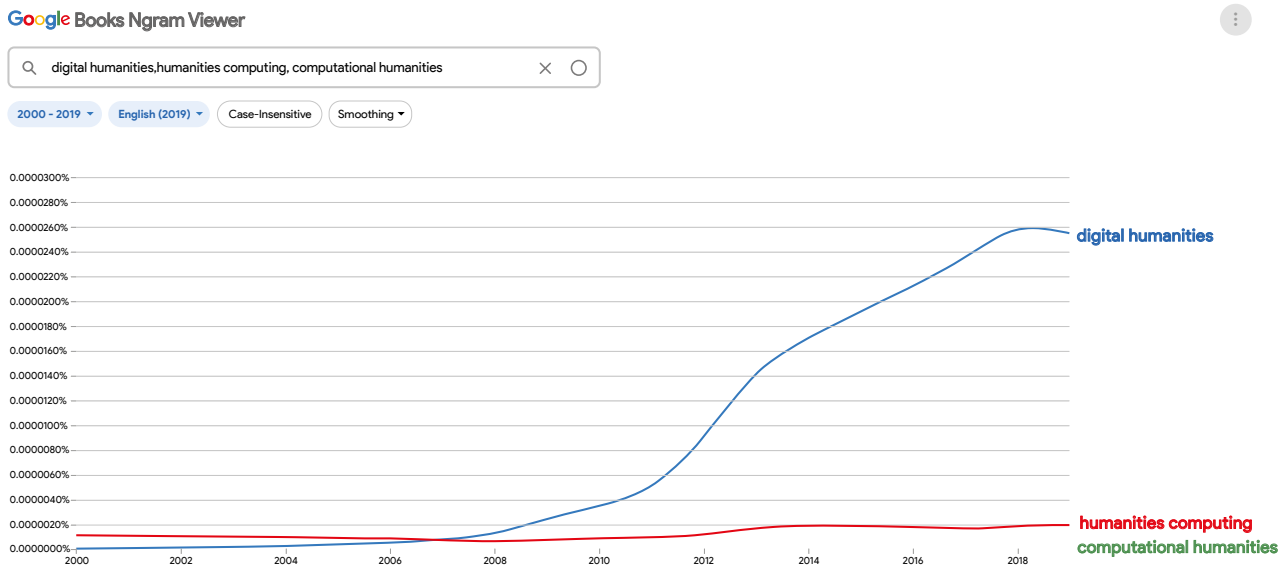
Although the integration of technology into the humanities as a field of study appears to be a new development, the intersection between the two fields has been bearing fruit for many years (Sahle, 2015). For several decades, the field of corpus linguistics has been producing increasingly large collections of texts with progressively detailed annotations. These corpora have become the basis for building dictionaries and language tools. In turn, these tools are now the forerunners in today's machine-learning applications, which can convert speech into written text and vice versa, translate from one language to another, and generate images from a description, etc. Digital information has become an integral part of fields including health, culture, bureaucracy, education, research, and industry.

As the interaction between the humanities and technology has evolved, so have the labels used to describe it. Before the year 2000, and in parallel with the development of corpus linguistics, there was talk in certain research areas regarding *humanities computing* or *computational philology*. After the year 2000, the term *digital humanities* became the most widely accepted label – one that became part of the names of national and international associations, projects, and publications. Some might argue that today it is difficult to avoid this new digital reality. The curious thing is that some of the people who have been most active in the digital humanities in recent years have now decided to abandon the adjective *digital* and switch to *computational*, in a move to show closeness to *computational linguistics*. This can be seen in the names used at conferences, such as *computational humanities* or *computational literary studies*.

«Before the year 2000, there was talk regarding *humanities computing* or *computational philology*. After the year 2000, the term *digital humanities* became the most widely accepted label»

Today, a person interested in the intersection of the humanities and technology might wonder whether to enter into the field of the digital humanities or

HOW TO CITE:
Calvo Tello, J. (2025). Where will the digital humanities be in 100 years?: The humanities as a hope for the digital. *Metode Science Studies Journal, 15*, 3–29. **https://doi.org/10.7203/metode.15.27672**



The labels that have been used to designate the interaction between the humanities and technology have changed over time, but after the year 2000 the term *digital humanities* became the most accepted label. Despite this, in the last few years some experts in the digital humanities have begun to use the term *computational humanities*.

SOURCE: Google Books Ngram Viewer

the computational humanities; whether they should learn the TEI (Text Encoding Initiative) format or rather JSON (JavaScript Object Notation); or whether we should be looking for a small corpus or a large dataset on which to train an artificial intelligence (AI) algorithm. Moreover, with the vast number of tools available and the speed of technological development, the pace of research is accelerating. A PhD usually takes at least three years, and most take more than four. But how can a researcher choose a technology to specialise in and believe it will still be relevant in five years' time?

From my current perspective, ChatGPT is the technology that is currently getting the most attention. However, by the time this article might be read in a few years' time, another more powerful tool with a better marketing campaign may have taken the media throne and ChatGPT could well be a distant memory. Indeed, a few years ago, BERT (Devlin et al., 2019), a natural language processing model trained by Google, looked like it would stand the test of time. Previously, methods such as convolutional neural networks were the technologies of choice (O'Shea & Nash, 2015). Before that, Topic Modelling (Blei, 2012) all but promised to extract semantic information from text without reading it. In 2014, Ted Underwood wrote his report on literary genre classification for the HathiTrust Center (Underwood, 2014). He pointed out that they had used a logistic regression algorithm and that more advanced methods did not seem to give better results. Ten years have now passed, and in that decade several generations of technology have emerged and, in some cases, even been surpassed themselves.

«With the vast number of tools available and the speed of technological development, the pace of research is accelerating»

THE PROBLEM: DIGITAL OBSOLESCENCE

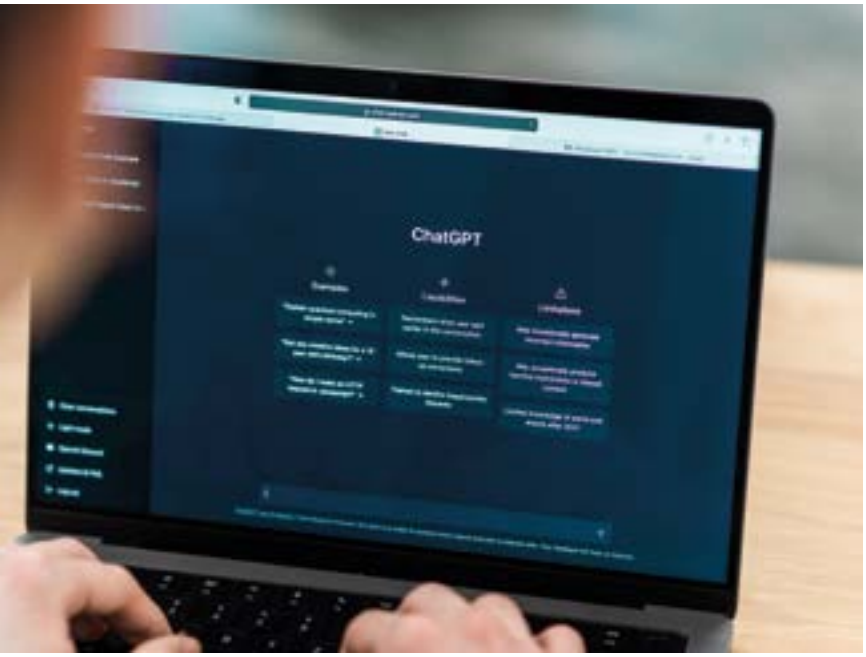
This accelerated process of technological development leads to paradoxes such as the one formulated by Rockwell in a talk shortly after the publication of *Hermenuetica* (Rockwell & Sinclair, 2016). Namely the fact that many digital humanities projects (or other areas in which technology and the humanities cross over) are working to develop tools that will be obsolete or no longer functioning by the time the project is completed. While this view may seem overly pessimistic, the fact remains that few digital assets survive more than five years. By the time they reach the age of ten, they are often seen as digital grandpas, respected, visited and quoted from time to time, but effectively retired. Rockwell himself

Image by frimufilms at Freepik

was involved in the development of the Voyant tool (Sinclair & Rockwell, 2016). Not only is this tool one of the applications that is still accessible years later, but in recent years it has evolved into Spyral,[1] an environment similar to that offered by Jupyter Notebooks.

Obsolescence seems to be an intrinsic feature of technology, and so there should be nothing to criticise. However, in many other areas of research, we find such impermanence unacceptable. Imagine if a library decided to discard books only ten years after they had been published; if our PhDs expired in a decade; an academic journal had to be discarded as obsolete after 10 years; or if we could not cite a printed edition of a literary text published more than a decade ago. These examples may seem absurd to us, but they reflect some of the more fragile aspects of the digital world in general, and the digital humanities in particular.

[1] https://voyant-tools.org/docs/#!/api



At the time of writing, ChatGPT is the technology that is receiving the most attention, but in the current context this may change in a very short time.

«Many digital humanities projects are working to develop tools that will be obsolete or no longer functioning by the time the project is completed

This is not just a problem of tools; it is a problem of data. Digital data must always be encoded in digital formats. These formats require tools to access, read, and display them. In other words, all digital data requires certain tools in order to be used. If the digital tools stop working, the digital data becomes obsolete, digital waste.

An example of this is Amazon's change in e-book format. Although various actors developed ePUB as the standard format for e-books (Garrish, 2011), Amazon decided to push its own formats: MOBI and AZW (McIlroy, 2012). This meant that if projects and publishers wanted their digital texts to be readable on Amazon's best-selling Kindle e-reader, they had to publish them in this proprietary format. In 2023, Amazon announced that some of the Kindle's features would no longer work with MOBI files (Mandal, 2023) and so, in a few years, Amazon may no longer provide support to open these files on its devices. It is even possible that no application will be able to open MOBI files. The result is that a project that has only retained editions in the MOBI format could find itself resigned to the digital graveyard.

This situation of potential infrastructure collapse also affects projects that have published their digital editions on their own websites. According to a recent survey, this option remains the most popular in the Spanish-speaking research community (Del Rio Riande & Allés-Torrent, 2023). Although the standard technologies behind websites (HTML, CSS, MySQL, JavaScript, and PHP, etc.) have a much broader development base than MOBI, these technologies evolve, functionalities are declared obsolete and no longer work, and server fees have to be paid annually, and so on. The flashier a site is, the more technologies we use to develop it, and the more innovative it is, the faster it will probably stop working. How many of these digital edition sites will still be available on the internet in 10, 20, 50, or 100 years?

FAIR AND REPOSITORIES: A PROBLEM SOLVED?

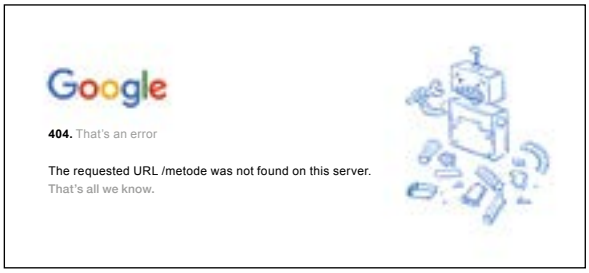
Partly to counter this situation of digital obsolescence, digital research has adopted the FAIR principles as one of its main guidelines in recent years (Wilkinson et al., 2016). FAIR stands for *findable, accessible, interoperable,* and *reusable*. Following these principles, researchers need to ask themselves questions and make decisions about how to make

their data more FAIR. Each of the components can be further broken down into a number of more specific principles, each of which can be applied to many aspects of data and related affairs. FAIR is understood as a set of principles rather than criteria, since in practice it is impossible to say whether a dataset is FAIR or not. Rather, we must argue whether or how a given dataset is more FAIR than another. Several publications in recent years have explicitly discussed the role of these criteria in the humanities, such as in the special issue of the journal *RIDE,* which reviews digital literature editions and corpora (Gengnagel et al., 2023).

Some of the FAIR criteria are more concerned with the way in which data are made available to the community. For example, data must be given a unique and persistent identifier (e.g., a Digital Object Identifier or DOI), the protocol by which the data are made available must be open and there must be a system for identifying users if necessary, or at least the metadata must be available in the long term. However, no one expects specific projects or individuals to manage their own DOIs or to worry about what kind of technical protocol to implement. It is taken for granted that projects will make use of existing infrastructures that provide solutions to these questions.

Nonetheless, exactly what kind of infrastructures should provide answers to the problems of DOIs, protocols, or long-term archiving? The answer lies in data repositories, i.e., platforms where researchers can publish their data and where the data is maintained in the long term. Examples of such repositories are Zenodo, the Consorcio Madroño data repository, TAPAS, TextGrid, or GAMs. In addition, these repositories should not only be funded by temporary research projects because otherwise they could suffer the same fate as the EVI-LINHD tool for editing and publishing texts (González-Blanco et al., 2017) that was abandoned after a few years when the funding of the project ended.

Is the problem of digital obsolescence solved with repositories? No, not at all. Firstly, because repositories today are only responsible for data for a period of ten years, which can be extended. Secondly, because we still have the problem of the obsolescence of other technologies associated with data. Let us look at a hypothetical example to get a better understanding: in 2023 we can publish a corpus of e-books in the MOBI format in a repository. We can expect a person to be able to download this corpus in 2033 but by then that person may not be able to view any of these documents. That is, repositories will hold data in obsolete formats



One of the problems facing the digital humanities is digital obsolescence. Few digital resources reach five years of life. This problem extends to the websites that host the projects. The flashier and more innovative our web portal is, the more likely it is to stop working in the short to medium term.

«All digital data requires certain tools in order to be used. If the digital tools stop working, the digital data becomes obsolete, digital waste»

and people will therefore have access to downloadable digital nonsense. The data is not completely lost, but it will have become useless.

Unfortunately, the fight against digital obsolescence is not solved simply by placing data in a repository. The data must also be in formats that comply with the FAIR principles. For the humanities in general, but more specifically for literary and linguistic studies, the TEI is one of the most relevant formats in this sense. Unfortunately, the adoption of TEI in the Spanish-speaking community has been truncated by a series of bad decisions made since the 2000s, which continue to burden the digital humanities community to this day (Allés-Torrent & del Rio Riande, 2019). In addition to format, several aspects of metadata are central to the FAIR principles: metadata must be standardised, using controlled vocabularies, classification systems or authority files, ideally all open and published online.

The future I outline may seem overstated, but the problem is more current than it seems. Since the mid-20th century, libraries have been working with a variety of formats and media beyond print or manuscripts. As an example, go to your trusted historical collections library, find someone who works with microfilm, and ask them if all of their microfilm holdings have been digitised. If they have not, ask them what happens to the data contained in those microfilms as the material deteriorates over time.

Up until now, the real and known risk of data loss in libraries has been relatively low, because microfilm represents a small percentage of the material in most

institutions. However, as I said at the beginning of this text, digital information is now a central part of our society. We can expect that technological obsolescence will affect more and more areas. According to Rockwell, we can also anticipate that the time between data generation and obsolescence will shorten. Thus, what was once a 50-year cycle for microfilm could be reduced to 10 or 20 years for digital formats. Is it acceptable for a person to lose their health records several times during their lifetime?

A CRITICAL PERSPECTIVE OF THE HUMANITIES AND QUANTITATIVE COMPUTATIONAL METHODS

In the title of this article, I promised a possible hope for this bleak digital outlook. So far, the digital humanities have focused more on applying digital techologies to the humanities. However, perhaps it is time to reorient and apply a humanities perspective to the digital world. After all, unlike many other scientific or engineering disciplines, most of the humanities work with historical materials. Some specific areas of the humanities (such as philological or historical editing, translation, and lexicography) are concerned with identifying, preparing, updating, and publishing historical materials for contemporary needs. Theoretical questioning and meta-reflection, so characteristic of the humanities, are aspects that can help the digital to become more sustainable over time.

For this humanistic critique of the digital to be understood by our colleagues in the engineering and research funding institutions, the humanities would do well to use more than their traditional theoretical critique alone. Otherwise the message will not be heard or understood by the audience they want to reach. Criticism should be underpinned by quantitative and computational methods that can be understood and that can convince researchers and decision-makers about the research. In other words, critique of digital solutions should be made by the digital humanities. Thus, I argue that the digital humanities should combine their enthusiasm for the digital (trying to convince colleagues in their departments that the digital humanities make sense) with a critique of aspects of the digital sphere that do not yet provide acceptable answers for the humanities.

Freepik

Until now, the real risk of data loss in libraries has been relatively minor, but digital data is now a central part of our society, and technological obsolescence can be expected to affect more and more areas.

«The digital humanities have focused more on applying digital techologies to the humanities. Perhaps it is time to apply a humanities perspective to the digital world»

Indeed, in a recent article my colleagues and I published a humanistic critique of digital elements using quantitative methods. We wanted to draw attention to the fact that the Integrated Authority File used in German-speaking countries (GND) and used in library cataloguing, has a (partly understandable) bias in favour of authors and works from the German-speaking area (Calvo Tello et al., 2023). More specifically, we suspected the GND would contain more authors and works from German-speaking countries than from those speaking other languages and that these data would be described with more information than entries from other languages. By comparing GNDs with other comparable resources (VIAF and Wikidata) and analysing several hundred literary works from the ELTeC corpus, published in 15 different European languages (Burnard et al., 2021; Schöch et al., 2021) we indeed found more German entries with more detail, although the number of authors from German-speaking or certain other European countries was similar. This paved the way for discussion based on quantitative results rather than theoretical arguments about possible information biases in libraries.

I mention the example of libraries because these institutions have a key role to play in this new digital phase. Libraries are among the research-related institutions where the historical gaze and issues of preservation and longevity are most present. Not surprisingly, in many cases libraries are behind research data repositories. As public services, libraries are central to ensuring that data can be stored and downloaded free of charge; in doing so, they not only maintain their social and democratic function, but also extend it into the digital domain. Their expertise and management of metadata in the catalogue and authority files can help researchers to improve the FAIR status of their research data, making it more discoverable, accessible, interoperable, and reusable. Therefore, the new role of libraries in the digital paradigm should also extend to conversion between formats. In the same way that libraries are responsible for repairing or replacing volumes when they are no longer usable, they will, at some point, also need to ask themselves whether they should convert certain digital formats into more current ones to preserve the data they contain for use in 20, 50, or 100 years' time.

«What will be the fate of each of our current resources? Which ones will disappear, which will be useless, and which will be useful?»

Image by rawpixel.com at Freepik

Let us return to the question I posed in the title of this article: what will the digital humanities be in 100 years' time? It is likely that in 10 years most young researchers using technology to analyse aspects of the humanities will no longer be using the label *digital humanities*. Twenty years from now, the digital humanities will likely be a thing of the past, overtaken by new labels and approaches that fit the times. But labels aside, in 50 or 100 years, what will be left of the digital work we have done in the digital humanities over the past few decades? None of the tools will work, and if any of them do, it will be in a very different way and with very different technology from what we use now.

As for the data (editions, databases, and corpora, etc.), many of them will be unusable: some will not even be visible and we will wonder if they ever existed; some we will discoverable but not openable; others will be viewable but we will not know how to use them; some will be in such outdated formats that we might either be too lazy to try to work with them or we will ask for new project grants to update the data to make them useful again.

However, some resources will have aged well and we will still be able to find them, open them, and work with them reasonably well. What will be the fate of each of our current resources? Which ones will disappear, which will be useless, and which will be useful? While we cannot fully predict which data or technologies will survive today's rapid evolution, we would do well to bring together the best of both worlds: the critical and historical perspective of the humanities and libraries could serve as the compass to point us in the right direction; the methods from the computational domain could move us from an explicitly qualitative to a quantitative scale. Future generations might thank us, and in a few years' time, so may we.

Libraries have a key role to play in the current digital phase. In many cases they support repositories of research data, and as public services they must ensure that these can be stored and downloaded.

«At some point, libraries also need to ask themselves whether they should convert certain digital formats into more current ones to preserve the data they contain»

REFERENCES

Allés-Torrent, S., & del Rio Riande, G. (2019). The switchover: Teaching and learning the text encoding initiative in Spanish. *Journal of the Text Encoding Initiative*, *12*. https://doi.org/10.4000/jtei.2994

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Burnard, L., Schöch, C., & Odebrecht, C. (2021). In search of comity: TEI for distant reading. *Journal of the Text Encoding Initiative*, *14*. https://doi.org/10.4000/jtei.3500

Calvo Tello, J., Rißler-Pipka, N., & Barth, F. (2023). GND und Normdaten für europäische Literatur? Personen und Werke in donen multilingualen Korpora von ELTeC. In A. Busch & P. Trilcke (Eds.), *Open Humanities, Open Culture, 2023. Konferenzabstracts* (p. 160–165). https://doi.org/10.5281/zenodo.7688631

Del Rio Riande, G., & Allés-Torrent, S. (2023). ¿Quién conforma la comunidad de la TEI en español? Análisis de los datos de una encuesta. *Journal of the Text Encoding Initiative*, *16*. https://doi.org/10.4000/jtei.4927

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv, 1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

Garrish, M. (2011). *What is EPUB 3?* O'Reilly Media. http://shop.oreilly.com/product/0636920022442.do

Gengnagel, T., Neuber, F., & Schulz, D. (2023). FAIR enough? Evaluating digital scholarly editions and the application of the FAIR data principles. *RIDE*, *16*. https://doi.org/10.18716/RIDE.A.16.0

González-Blanco, E., Cantón, C. M., del Rio Riande, G., Ros, S., Pastor, R., Robles-Gómez, A., Caminero, A., Díez Platas, M. L., del Olmo, Á., & Urízar, M. (2017). EVI-LINHD, a virtual research environment for the Spanish-speaking community. *Digital Scholarship in the Humanities*, *32*(suppl_2), ii171–ii178. https://doi.org/10.1093/llc/fqx025

Mandal, S. (2023, 2 February). Amazon's send to Kindle feature still supports sending MOBI file. *Good E-Reader*. https://goodereader.com/blog/kindle/amazons-send-to-kindle-feature-still-supports-sending-mobi-file

McIlroy, T. (2012). Ebook formats are a mess – here's why. *Learned Publishing*, *25*(4), 247–250.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv, 1511.08458. https://doi.org/10.48550/arXiv.1511.08458

Rockwell, G., & Sinclair, S. (2016). *Hermeneutica: Computer-assisted interpretation in the humanities*. The MIT Press.

Sahle, P. (2015). Digital humanities? Gibt's doch gar nicht! *ZfdG*. https://doi.org/10.17175/sb001_004

Schöch, C., Erjavec, T., Patras, R., & Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and perspectives. *Modern Languages Open*, *1*. https://doi.org/10.3828/mlo.v0i0.364

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools* [Software]. http://voyant-tools.org/

Underwood, T. (2014). *Understanding genre in a collection of a million volumes, Interim Report*. https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*. https://doi.org/10.1038/sdata.2016.18

JOSÉ CALVO TELLO. Subject librarian and researcher at the State and University Library Göttingen, Georg-August-Universität Göttingen (Germany). He received his doctorate from the University of Würzburg in 2020 with a thesis entitled *The novel in the Spanish Silver Age: A digital analysis of genre using machine learning* (transcript, 2021). His interests focus on the development and application of quantitative methods to Romance literature and library collections and the preservation of data from literature studies. ✉ calvotello@sub.uni-goettingen.de