Visualizing the Catalogs of Digital Editions

MICHAEL KURZMEIER, JAMES O'SULLIVAN, MICHAEL PIDD, ORLA MURPHY, AND BRIDGETTE WESSELS

Abstract: This article provides a data-driven overview of the developments in the field of digital scholarly editing. It surveys and evaluates the available data source on digital scholarly editions and provides longitudinal analysis of changes in number of projects, geographic distribution, licensing, interfaces, and preservation. Digital scholarly editions (DSEs) are essential to arts and humanities research but also to society and culture at large. They are the primary instrument through which textual and cultural heritage, expert knowledge, and public understanding are negotiated. Their comparatively long history makes them especially suited for a diachronic approach, describing their change over time. While digital editions can vary greatly in scope and lifespan, a quantitative analysis of two of the most comprehensive data sources on digital editions can produce data-based insight into the developments within the field over time. Exploring this history and at the same time assessing the available metadata on DSEs is the aim of this article. It presents the state of the two most comprehensive available sources on digital editions and details the methodology and visualization process undertaken. In its analysis, it is a quantitative approach to DSEs as well as a critique of the available data sources on editions.

Keywords: Digital scholarly editing, digital scholarly edition, Preservation, TEI

Funding statement: This research was funded by the Irish Research Council and the UKRI-AHRC under the UK-Ireland Collaboration in the Digital Humanities Research Grants (IRC/W001489/1 and AH/W001489/1).

Data statement: All data and all code used in this research is freely available; see the project repository.

Declaration of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Table of contents

1. Introduction

Digital scholarly editions (DSEs) are essential to arts and humanities research but also to society and culture at large. They are a primary instrument through which textual and cultural heritage, expert knowledge, and public understanding are negotiated. Scholarly editions make documentary materials reliable, adhering to established editorial standards and rigorous intellectual practices to ensure that cultural artifacts are compiled and represented together with critical annotation and other forms of editorial enrichment.

DSEs are among the earliest projects in a field that used to be described as humanities computing and is currently mostly referred to as digital humanities (Dalbello 2011). This makes them especially suited for a diachronic approach, describing their change over time. While digital editions can vary greatly in scope and lifespan, a quantitative analysis of two of the most comprehensive data sources on digital editions can produce data-based insight into the developments within the field over time. This longitudinal approach to DSEs also means that historically successful formats for editions will shape the dataset and respective metadata.[1] In practice, this means that most of the editions in the data sources will be based on or modeled after critical editions of text.

1. The notion of success in relation to digital edition projects is, of course, highly subjective. For the context of the methodology employed in this article, *success* is defined in evolutionary terms: a feature or standard is successful if it perpetuates itself. Also see Unsworth (1998).

Digital editions are positioned between drawing from archived material and being an archive themselves (Dillen 2019, 266). In addition, digital editions are web resources in need of archiving, lest they fall subject to link rot and very soon disappear from the web either for the lack of a persistent identifier or lack of maintenance. A 2010 report on digital preservation of cultural heritage objects finds that obsolescence in hard- and software is a likely reason for information to become unavailable (Kirschenbaum et al. 2010, 17). James Cummings reminds researchers that "we should always plan for events that affect the sustainability of digital research projects. Some of these events are obviously unpredictable, while others are of our own making" (2023). For digital editions past and present, two main data sources are available. Patrick Sahle lists around 800 editions in a curated catalog (Sahle 2020), and the Catalogue of Digital Editions (Dig-Ed-Cat) features about 320 digital editions in a database (Franzini 2022). Both sources have different criteria for inclusion, overlap in content, and differ in granularity, yet these are the sources from which a quantitative history of DSEs will mostly draw. Analysis of these sources will present them in their scope, aim, and usability for research while highlighting underrepresented areas of data collection on DSEs.

DSEs in the context of this article are best defined using Patrick Sahle's definition of editions as the "critical representation of historic documents" (2016). This definition is best suited for the editions described in both data sources, as Sahle compiled the first list of digital editions and the Dig-Ed-Cat explicitly refers to Sahle's work as the basis of its data collection. These data sources can form the basis of a response to Joris van Zundert, who calls on theorists and practitioners to "intensify the methodological discourse" necessary to "implement a form of hypertext that truly represents textual fluidity and text relations in a scholarly viable and computational tractable manner" (2016, 106). "Without that dialogue," he warns, "we relegate the raison d'être for the digital scholarly edition to that of a mere medium shift, we limit its expressiveness to that of print text, and we fail to explore the computational potential for digital text representation, analysis and interaction." This dialogue has begun in earnest (Driscoll and Pierazzo 2016; Boot et al. 2017), but a previous survey on the expectations and use of digital editions found that user needs are seldom satisfied by such resources (Franzini, Terras, and Mahony 2019). This dialogue can be augmented by the availability of data-driven insight into the development of DSEs over time, as well as an assessment of the scope and potential use of data sources for research on digital editions.

Digital editions have a long history of adapting to new challenges and possibilities presented by both changes in computational methods and research policy. While earlier DSEs were distributed through physical media, such as a CD-ROMs sold through a publisher, digital platforms and browser-based access have become the standard practice for digital scholarly editions. With recent changes in EU funding directives, the majority of recent digital editions are based on open access models, further increasing accessibility of digital editions. Comparative legislation in the United States and United

Kingdom also give preference to open access models. Similarly, digital platforms providing metadata about digital editions can extend the accessibility of scholarly editing by providing the data from which to construct a history of digital scholarly editing and editions. Minimal computing editions are an approach to deliberately reduce the technical complexity of digital editions to increase accessibility and longevity (Risam and Gil 2022; Siddiqui 2022). Despite these developments, no widely used archiving solution for DSEs exists. Extending the reach of the collected data in both sources, this article also provides new data on the long-term availability of digital editions.

Exploring this history and at the same time assessing the available metadata on DSEs is the aim of this article. It presents the state of the two most comprehensive available sources on digital editions and details the methodology and visualization process undertaken. In its analysis, it is simultaneously a quantitative approach to DSEs and a critique of the available data sources on editions. A discourse such as van Zundert called for must be systematic and based on a solid data corpus regarding the history and development of the field. By providing an overview of the available data and an estimation of DSEs already lost, this article aims to provide part of the foundation for a data-driven discourse on DSEs. This article is part of the larger C21 Editions project, a three-year international collaboration jointly funded by the Arts & Humanities Research Council (AH/W001489/1) and the Irish Research Council (IRC/W001489/1).

2. Methodology

2.1 Data Sources on DSEs

Metadata on DSEs is primarily available through the Catalogue of Digital Editions (Dig-Ed-Cat). It incorporates other sources such as Patrick Sahle's list of DSEs and provides new data fields. This makes the Dig-Ed-Cat a key resource for a quantitative analysis of DSEs over time, with the project itself providing multiple ways to access and query the data. The Dig-Ed-Cat also fulfills a second, but equally important function: through syndication with the German Datenbank-Infosystem (Database information system; DBIS), DSEs listed in the Dig-Ed-Cat are discoverable resources for subscribing libraries. This greatly adds to the findability and impact of listed DSEs and is an important step in the integration of DSEs with other academic resources.

The common source of both catalogs can be traced back to a list of digital editions Sahle compiled in 1997.[2] Continuously updating the list, Sahle would expand

2. See Sahle's "About" for version 3.0 of his list of DSEs, https://v3.digitale-edition.de/vlet-about.html: "I've been interested in digital scholarly editing roughly since 1994. I still try to keep an eye on the ongoing developments in this area and

the catalog to a list of 714 currently available editions. In a comment on the collection, Sahle explains his motivation for compiling the list:

> Since most digital editions don't really fit into the traditional bibliographic model, usually there are no bibliographic library records available for them. We try to create bibliographic-like data as far as possible by collecting names of general editors, places of (virtual) publishing, publishing institutions, years of publishing and ISBN-numbers and other identifiers (where available) to help make these editions identifiable and referenceable. Whenever possible, we use snippets of self-description from the web pages of the editions as comments. Otherwise, there are edition descriptions from my (Patrick's) point of view. This is particularly the case for the oldest entries. Remember: some entries go back to the late 90s, early 00s and the concept of self description by quotation only started in 2008. (2020)

This aim of increasing discoverability of digital editions in the light of poor bibliographic records for such works (expressed in both the 2008 and 2020 version of the catalog) is again found in the Dig-Ed-Cat's syndication of editions to increase discoverability through library catalogs. Sahle's list is compiled by himself and updated irregularly. Regarding data bias, Sahle explains, "What I see is what you get. I am currently professor for Digital Humanities, affiliated to a history department at Wuppertal University, Germany. I have a background in (medieval) history and historical and humanities methodologies at large. Therefore, German editions will inevitably be overrepresented as well as historical editions, editions from Europe, and editions documented in Western languages. You can improve this situation by indicating underrepresented types of editions" (2020). This data bias is also found in the Dig-Ed-Cat, for which the authors explain that a lack of Asian or African digital editions in the catalog is due to the project team not being able to evaluate these editions. Instead, users are invited to submit entries on editions from the Global South.[3] With both data sources, it must be considered that the data held is largely focused on European projects.

While both Sahle's catalog and the Dig-Ed-Cat are generally open to additions, Dig-Ed-Cat is designed to be supplied by the DSE community. The project allows new project entries to be submitted through GitHub or a Google Form. These submissions are then checked by the project team and added to the database. While this approach allows for a greater number of projects to be featured, it also presents a potential conflict

I continuously collect hints on digital editions. This list is replacing my Virtual Library Page on '(Digitale) Editionstechnik' from the year 2000 which I retrospectively would call Version 2.0. There were even earlier lists from 1998 and 1997 which I now call V1.0 and V0.8 respectively."

3. See the FAQ section of the Dig-Ed-Cat, https://dig-ed-cat.acdh.oeaw.ac.at/faq/.

with the proposed granularity of Dig-Ed-Cat, as (especially when using the Google Form) data submitted by users is not always complete or correctly structured. Whereas the Dig-Ed-Cat maintains a list of projects considered not suitable for inclusion in the catalog (the list only names the projects, not time of submission or reasoning), Sahle's catalog does not feature a list of surveyed, but not-included projects. It is further worth noting that Sahle's catalog is approximately twice as large as the Dig-Ed-Cat, yet no statement is available as to the inclusion criteria Dig-Ed-Cat applied to Sahle's data. It might be that the crowd-curated approach of the Dig-Ed-Cat has not produced the same level of granularity as Sahle's curated list, or it might simply be the case that inactive or long-gone edition projects are not in scope for the Dig-Ed-Cat.

Both data sources generally focus on open access publications, although this is not stated in their collection statements. In practice, this creates a bias toward editions published through funded research projects, as these projects usually result in open access publications. Paid access publications are generally underrepresented, so that the data sources may be more accurately described as *DSEs produced as part of funded research projects*.

2.2 Structure

As mentioned, both data sources are different in structure and granularity. Sahle describes his catalog not using a controlled vocabulary and using few categories (title, subject area, language, material, period). Additionally, each entry has a description/comment and a link to the edition or publisher. Sahle deliberately does not remove inactive projects (though it is not clear if and when broken links are removed or updated).[4]

The Dig-Ed-Cat self-describes as a data agglomerator that again is connected to the Linked Open Data cloud to promote discoverability. Consequently, Dig-Ed-Cat features a much more granular approach to DSEs, including some fields with linked data vocabularies.[5] Dig-Ed-Cat is designed as a participatory project, and users can submit new entries through the project's GitHub page. One of the data fields in the Dig-Ed-Cat is availability, indicating that inactive projects remain part of the data; however, it seems that this value does not update automatically based on, say, the HTTP status of the URL. As part of the granular approach and voluntary contributions from the DSE community, some fields concerning funding source and amount per project are seldom filled out. The

4. Explained in the "About" section of Version 4.0; see Sahle (2020).

5. "The data in the *Catalogue* is being hooked up to the Linked Open Data (LOD) cloud to increase discoverability while supporting semantic integration and knowledge sharing. The linked data vocabularies currently connected to the *Catalogue* are W3C Basic Geo, FOAF, GeoNames, Data Catalog and Dublin Core Metadata. Institution data also links to DNB (Deutsche Nationalbibliothek) Catalogue IDs" (Andorfer, Zaytseva, and Franzini 2018).

analysis in this article will largely be based on the granular data available through the Dig-Ed-Cat, especially with regard to longitudinal developments in the field.

2.3 Data Access

As a data agglomerator and data source, Dig-Ed-Cat features convenient access to underlying data. While the project home page features extensive data visualizations and a SPARQL end point, it also provides open access to the underlying data that can be accessed in. csv format.[6] Sahle's catalog—referring here to version 4.0, released in 2020—also has made available its data source.[7] The data is structured in XML-TEI and can be translated into a DataFrame or. csv file for a quantitative analysis.[8] Before version 4.0, the website itself could be scraped for edition data, although this includes somewhat more effort especially in regard to data cleaning.

2.4 Analysis

With the data available, the main objective of the analysis is to produce a longitudinal analysis of DSEs within the scope and constraints of the respective data sources. In most cases, the more granular Dig-Ed-Cat is the preferred source (especially since Dig-Ed-Cat provides an end date field indicating the actual or envisioned end of the active development), while in some cases the larger data source provided by Sahle yields more conclusive results. Acknowledging the bias and limitations in both sources, the aim of this analysis as a whole is to understand long-term developments within the field of DSEs and at the same time show the value of comprehensive data sources on digital editions for research purposes.

3. Key Findings

3.1 Overview

To gain a first overview of the dataset, the first step in visualizing is to show editions over time. For this scenario, the data extracted from Sahle's catalog can be used to show the number of editions over time. This results in the visualization provided in Figure 1.

6. SPARQL is a semantic query language for databases, allowing users to run custom queries on the Dig-Ed-Cat database.
7. https://git.uni-wuppertal.de/dhsfu/sde-catalog
8. For a detailed description of this process, see the section on availability.

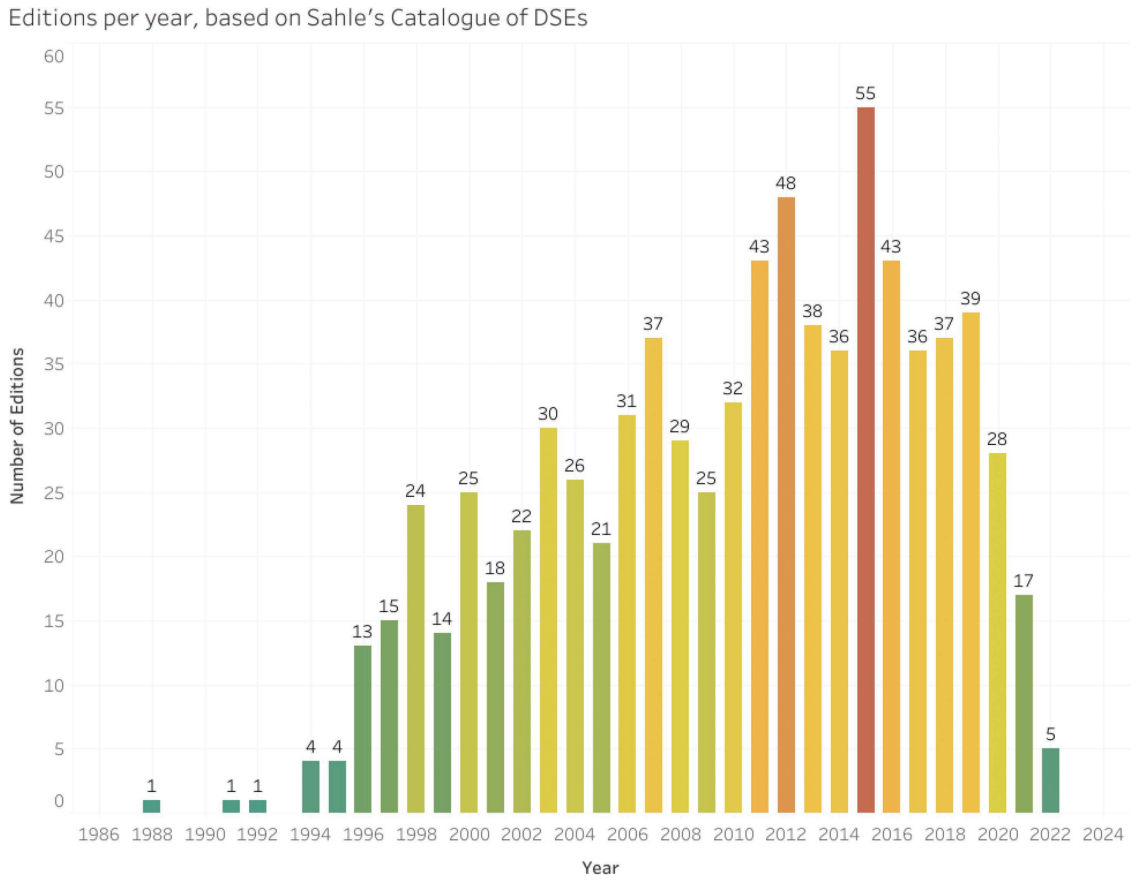

Figure 1. Editions per year, based on Sahle's catalog of DSEs

Note that Sahle's catalog uses multiple dates for editions, including first publication, second editions, and relaunches. The data used here is the first publication date. Figure 1 shows that the oldest edition in the data source goes back to 1988, while the latest entry is from 2022. The number of editions increases over time until 2015, then sharply declines. This may be due to the fact that editions only become part of the catalog after publication and works in progress are less likely to be featured. Another possible explanation is a reduced data collection activity.

Historicizing editions, Amy Earhart describes a general distinction between *like print* and *better-than-print editions*: "It is the digital edition's apparent similarity to print that reassuringly runs through early digital editorial work; the hallmark of the early digital edition is the sense of stability and the reassurance of forms that look like a print text. And while those forms might remain stable, the digital allows for the creation of better-than-print editions" (2012, 22). While these changes have no definitive timeline, it is reasonable to locate the distinction between early and late digital editions in the first decade of the twenty-first century. Together with this changed approach came a proliferation of tools and licensing models that led to a proliferation of digital editions.

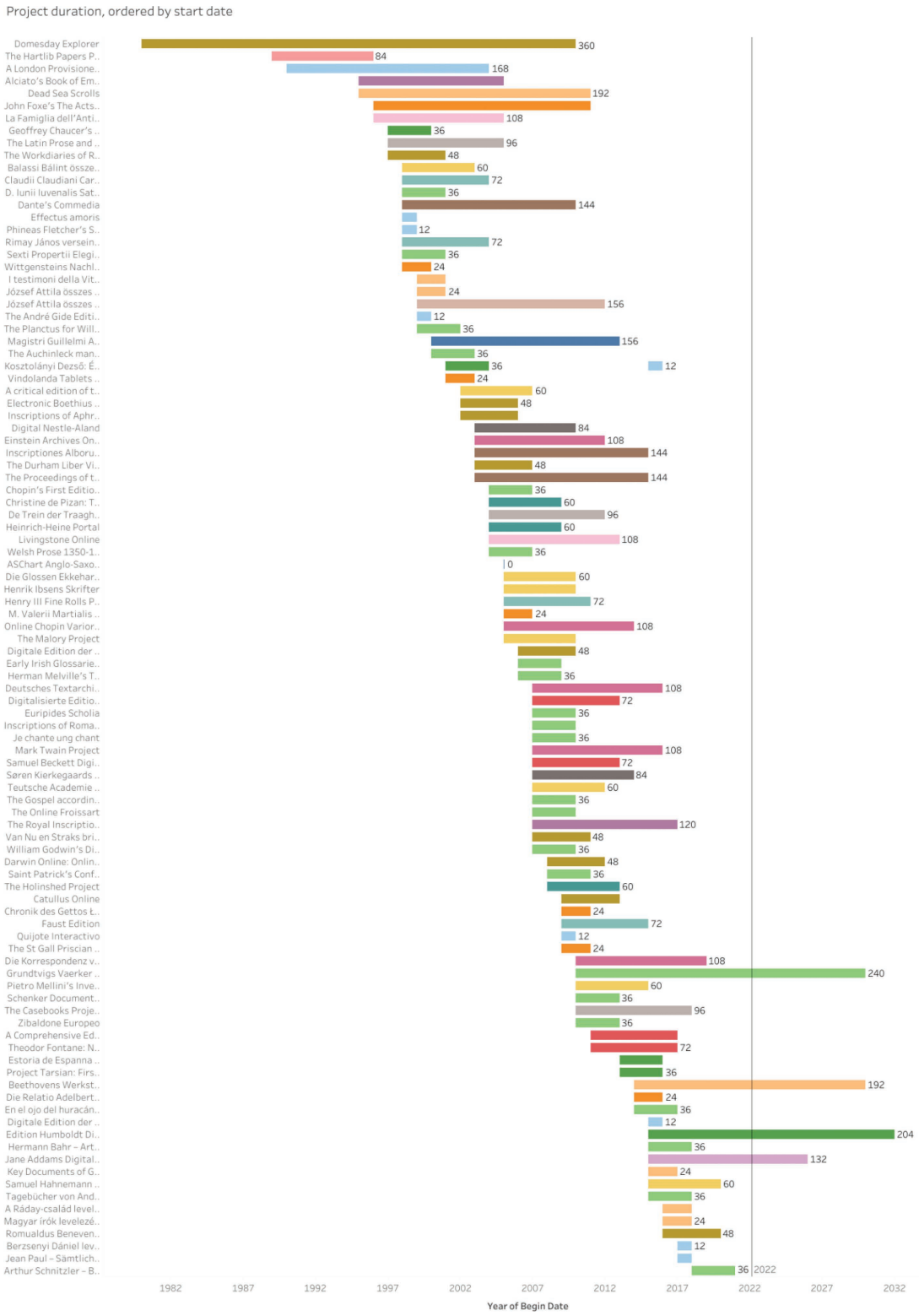

Figure 2. Project duration by start date

To get a better understanding of the works in progress and the average lifespan of a digital edition, Dig-Ed-Cat provides a value for the end date (usually this indicates the official end of the project and funding). Complementing the two sources, it is possible to plot individual editions ordered by their start date, as is done in Figure 2.

Since the Dig-Ed-Cat features different date formats (likely due to non-standardized user inputs), and to help the readability of Figure 2, start dates are reduced to the year, while the difference between start and end dates is calculated in months. This difference then allows us to assign the color and size. Noteworthy findings include one edition project with a reported duration of zero, one edition with two periods of activity, and a number of long-term edition projects with future end dates. Since not all entries in Dig-Ed-Cat have an end date value, this visualization only includes editions with both start and end dates. Similar to the previous visualization (Figure 1), it seems likely that ongoing projects are less likely to have an end date, skewing the results accordingly. Note that not all project entries have an end date value; this might mean either they are still ongoing or the user left the field blank when submitting the data.

The duration of edition projects is a useful value that can be plotted over time to show the average project duration per year, as is done in Figure 3. This graph again must be considered in the light of the scarcity of digital editions pre-1995. The few very old

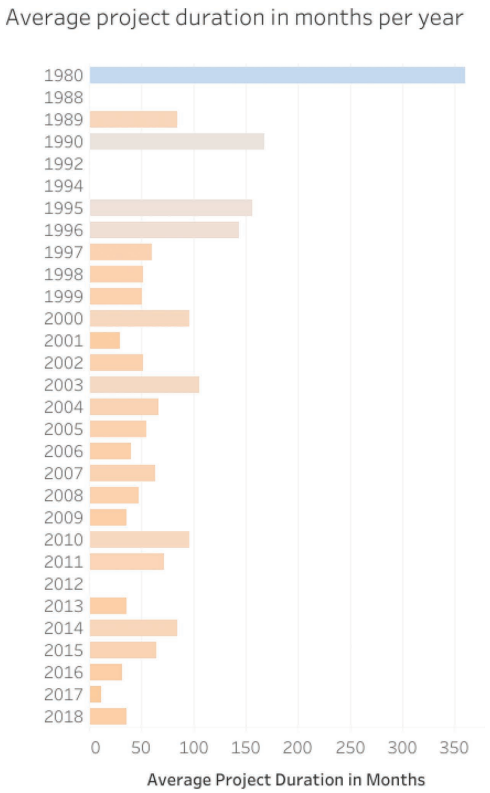

Figure 3. Average project duration in months per year

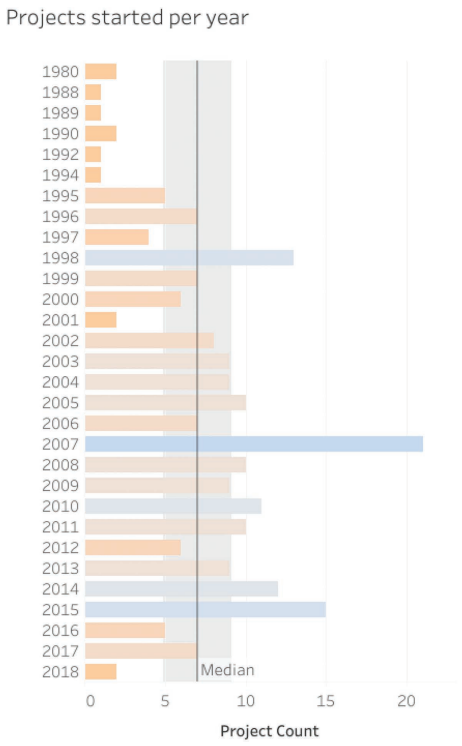

Figure 4. Number of projects per year

editions featured in Dig-Ed-Cat disproportionately influence the average project duration for these years. Nevertheless, it can be shown that the average project duration is thirty-six months, with a tendency to decline past 2010. The reason for this is speculative but may be a combination of the availability of comprehensive tools to produce editions (mostly for textual editions), without the need to develop a custom solution from scratch, and also the fact that as digital editions become more widely accepted as scholarly research projects, editions adapt to funding cycles and aim for results achievable within these cycles.

Using the start date value, it is also possible to plot the number of projects started per year, as is done in Figure 4. This is helpful since it gives a better chance to also include ongoing projects that might not have an end date yet. This graph then is similar to the first visualization (Figure 1) and shows a general upward trend in the number of projects started per year. The average number of projects per year is around seven to eight. The drop in numbers after 2015 is likely due to similar reasons as described in the first visualization: unfinished projects are less likely to report or be reported to the Dig-Ed-Cat. Whether these figures imply an actual decline in the number of projects is unclear; as mentioned before, there is reason to assume a delay between the official project start and projects being first made available to the public (which in many cases would be when they are submitted to either data source).

3.2 Geography & Institutions

As both data sources mention the Global North bias in their data, another interesting approach to the data is to visualize the location and languages of digital editions. The Dig-Ed-Cat features coordinates for editions, which easily lets us plot their location on a map (Figure 5).

The map in Figure 5 confirms the concentration of DSEs in Europe and North America. Only four editions are located outside the Global North, with none at all being located in Asia or South America. Among the locations, the United Kingdom, United States, Germany, and Italy are the countries with the most editions. Such a result is hardly surprising given that both sources explicitly express their awareness of this bias in their documentation, yet it is a stark reminder that access to tools essential for the study of cultural heritage is subject to a strong divide between countries of the Global North and those of the Global South. Despite the mentioned proliferation of tools, the focus on the Global North remains. This may be traced back to a lack of funding as well as a lack of source material: "In academia, libraries, and the cultural heritage sector, there is simply not enough funding to fully redress the inequalities in the digital cultural record. Furthermore, the cultural record itself is disrupted by voices that were

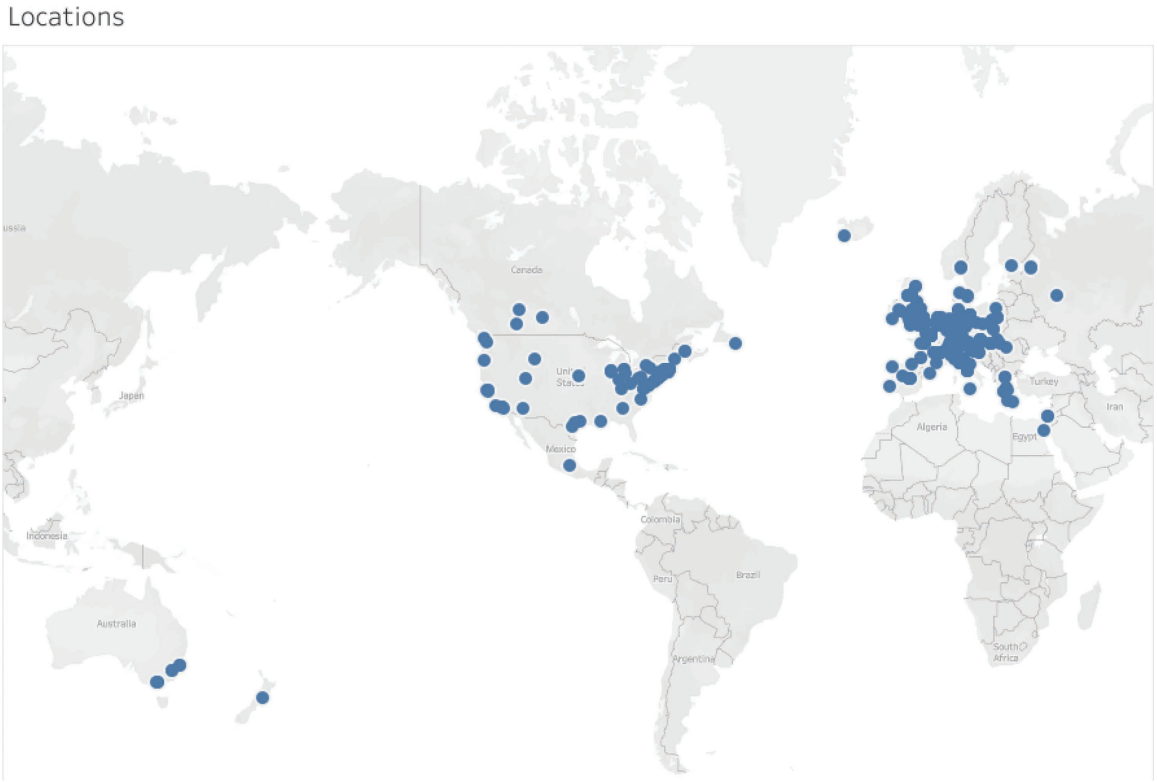

Figure 5. Location of institutions represented in Dig-Ed-Cat

never recorded, for which we will never be able to create digital representation" (Risam 2019, 19).

This is, of course, a critique aimed at the larger academic and cultural field and not specific to either of the data sources. Regardless, these inequalities are confirmed when looking not only at the institutions represented in the data sources but also at the project languages among DSEs. Dig-Ed-Cat allows combinations of languages within one field, and thus bilingual editions are less likely to rank highly. The language selection refers to the source material of the edition, not the interface (although it is unclear if all entries are correctly labeled).

The most widely used language is English, followed by Latin, German, Hungarian, Italian, and Spanish (Figure 6). This further confirms a bias toward North American and European institutions already seen in the previous visualization (Figure 5). This Western-centeredness is addressed by the authors in the FAQ section of the Dig-Ed-Cat:

> **Q. Why are there no Asian and/or African digital editions in the *Catalogue*?**
> Because the team cannot read those languages and we hesitate to rely on browser translation plugins as any machine translation errors would go unnoticed and possibly lead to incorrect cataloguing. We're eager to fill this gap but need help from users to do so. (Franzini 2022)

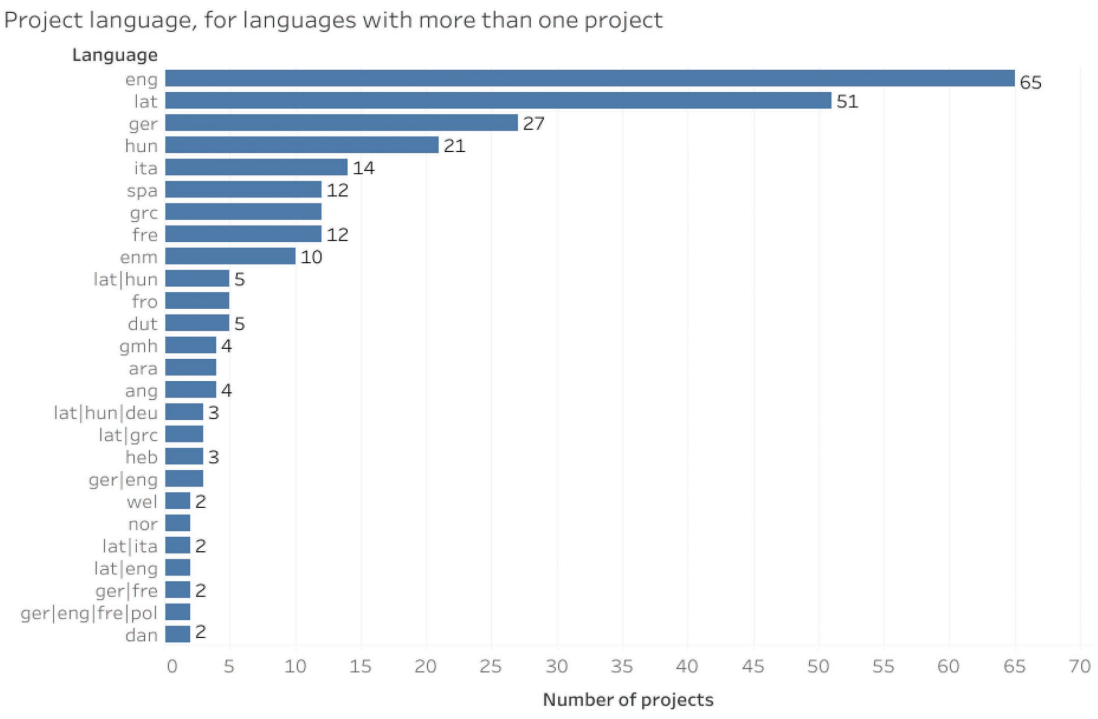

Figure 6. Primary material languages

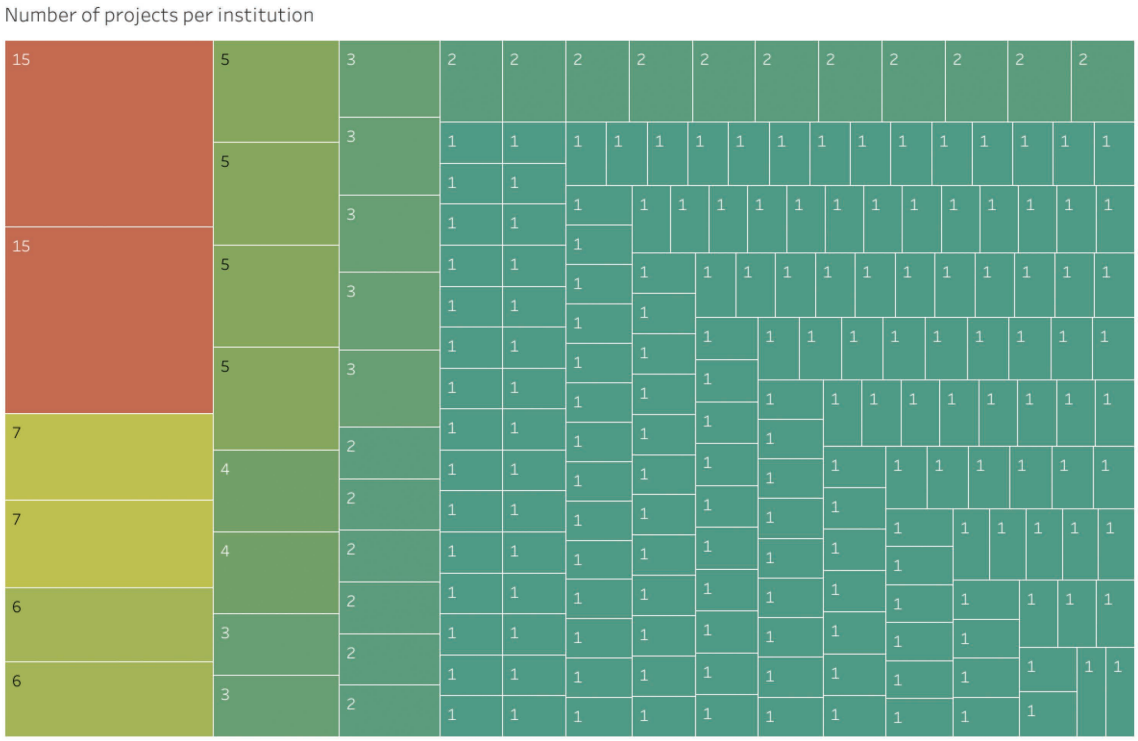

Figure 7. Number of projects per institution

What is of further interest is the level of centralization between institutions. While in Figure 5 each dot represents an institution, this does not account for the number of projects housed at that institution. The map reveals a distribution within Europe and North America but does not allow us to make clear statements about the number of projects per institution. Regarding the data source, Dig-Ed-Cat features institutions but also allows for a combination of institutions. Furthermore, editions might not be housed at academic institutions at all but at private or public institutions. Plotting the number of projects per institution produces the graph in Figure 7.

Because of the inconclusive data, the visualization in Figure 7 is to be taken as indicative only. Rather than rating institutions against one another, the point here is that the data on digital editions—as Western-centered as it is—is rather evenly distributed with only ten institutions having five or more projects on record. This may indicate that DSEs are more projects of individual scholars than institutions. The data in this case is hard to check for accuracy, with potential error sources being collaborations between institutions (University College Cork and University of Sheffield would, for example, count as one new institution rather than two).

This positive result, however, is most likely due to incomplete data. For example, Oxford University Press's *Oxford Scholarly Editions* alone hosts over 1,750 DSEs. As

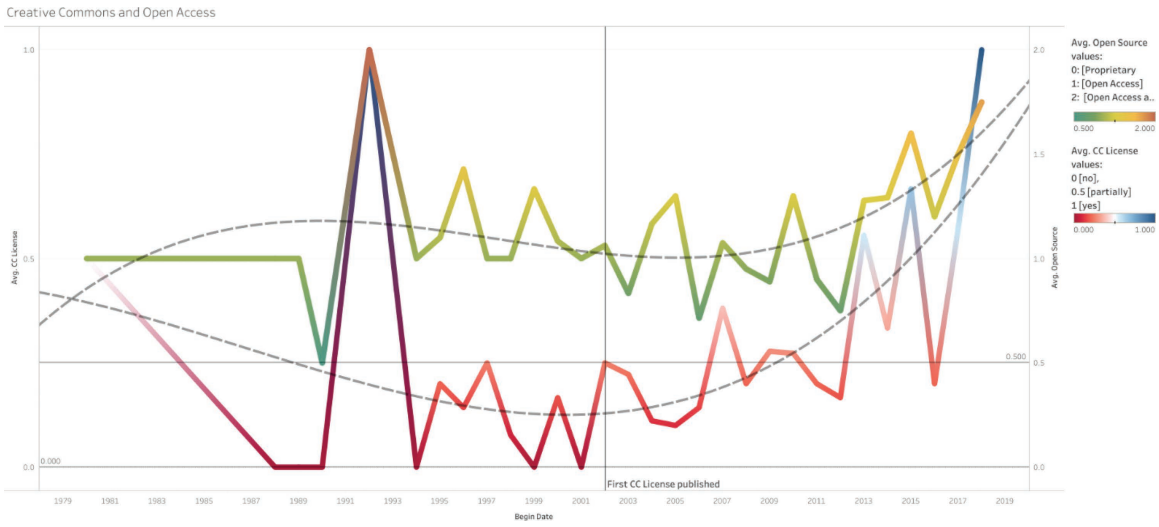

Figure 8. Open source licensing and open access models in DSEs over time

mentioned before, closed access publications are largely absent in both data sources, and thus the data available describes a specialized subset of DSEs—those created within institutions through research funding.

3.3 Licensing & Access

Even with the mentioned bias in the data sources, different models of licensing and regulating access to DSEs still exist. Dig-Ed-Cat records both the assigned license and the access model for editions. Using this data, it becomes possible to plot both in one graph (see Figure 8).

What we can see in Figure 8 is that both use of open source licensing and open access licenses increased in the early 2000s. It may be speculated what caused this trend, as it predates changes in funding regulations (in the European Union). Likely causes are a general move toward *open science* and a proliferation of open source tools that enabled smaller teams to produce editions with little to no financial requirements.

This figure also indicates inaccuracies within the data source: the first set of Creative Commons (CC) licenses was released in December of 2002.[9] Possible explanations include that CC licensing is understood *pars pro toto* as a description of permissive licensing in general (see a brief overview of free and open source licenses) or that editions changed their licensing model at some point after December 2002.

9. See https://web.archive.org/web/20091103232523/http://creativecommons.org/about/history.

Figure 9. Usage of XML-TEI over time

Further noteworthy here is the life cycle of such policy changes: while 2005 sees open source become the dominant (>50%) practice, the same is only true for CC licensing from 2015 on. With recent change in European Union funding directives (open access as the defined standard for research outputs where possible), these trends are likely to be sustained.[10]

3.4 The Text Encoding Initiative

When speaking of open tools to benefit the field of digital editions, the key tool that had a measurable impact on the field is the Text Encoding Initiative (TEI) Guidelines. Dig-Ed-Cat measures the transcription method on a scale from 0 (no XML and no TEI) to 0.5 (XML but no TEI) to 1 (XML-TEI). This allows us to plot the average value for this field over time (see Figure 9). From its official introduction in 1990, it took XML-TEI thirty-one years to establish itself as the dominant (used in >50% of editions) practice for source material encoding in DSEs. While its adaptation varied between 1987 and 2002, adaptation steadily increased from thereon. This data shows

10. See the EU Grant AGA 2021–2027 (European Commission 2021).

both the timescale of introducing a standard in encoding as well as the importance of XML-TEI. XML has been the dominant practice since 2002, and XML-TEI has been the dominant practice since 2012. This again relates to the drop in centralization and the increase in open source licensing in DSEs. Relating the last three quite different approaches to the question of how tools and service models impact the design of digital editions, we can now with some certainty say that the data shows a relationship between the use of open source tools and open access models and the proliferation of editions.

3.5 Interfaces and Interoperability

The question in this section relates to the possible different ways of accessing data in a digital edition. Two fields in Dig-Ed-Cat are potentially relevant in this regard. One measures the existence of application programming interface (API) access to the edition, and the other records the availability of a print-friendly view. The documentation indicates that both are binary questions, with the first one not specifying any details about the level of API access and the latter not recording details about the print function. While print function and API access at first glance appear to be very different things, in the context of access, they perform similar functions of helping to export data from the edition to use it in other contexts. While the general move toward open data in digital editions means that increasingly the underlying data is available, this does not equal easy access or export as is provided by the two functions described here. Figure 10 shows that API access is a rare function in DSEs and is limited to a few, more recent editions. The ability to easily print sections (pages) of editions has been a consistent feature that is used in more than half of editions before 2010 and approximately one-third of editions after 2010. However, Figure 8 has shown that editions are increasingly likely to provide open access to their underlying data, which is what users are primarily interested in.

3.6 Availability

For edition availability—that is, the assessment of how long an edition will be accessible after launch—both sources offer data. First, Dig-Ed-Cat features a value for *current availability*. This asks the submitting author if the edition is currently available. As Dig-Ed-Cat is an actively maintained resource, it seems that these checks are also performed periodically to identify broken links. The results of simply aggregating that data and comparing it against a check of all project URLs are shown in Table 1.

These preliminary results seem to confirm that loss happens among digital editions. The difference between the two checks might be due to the time difference and likely

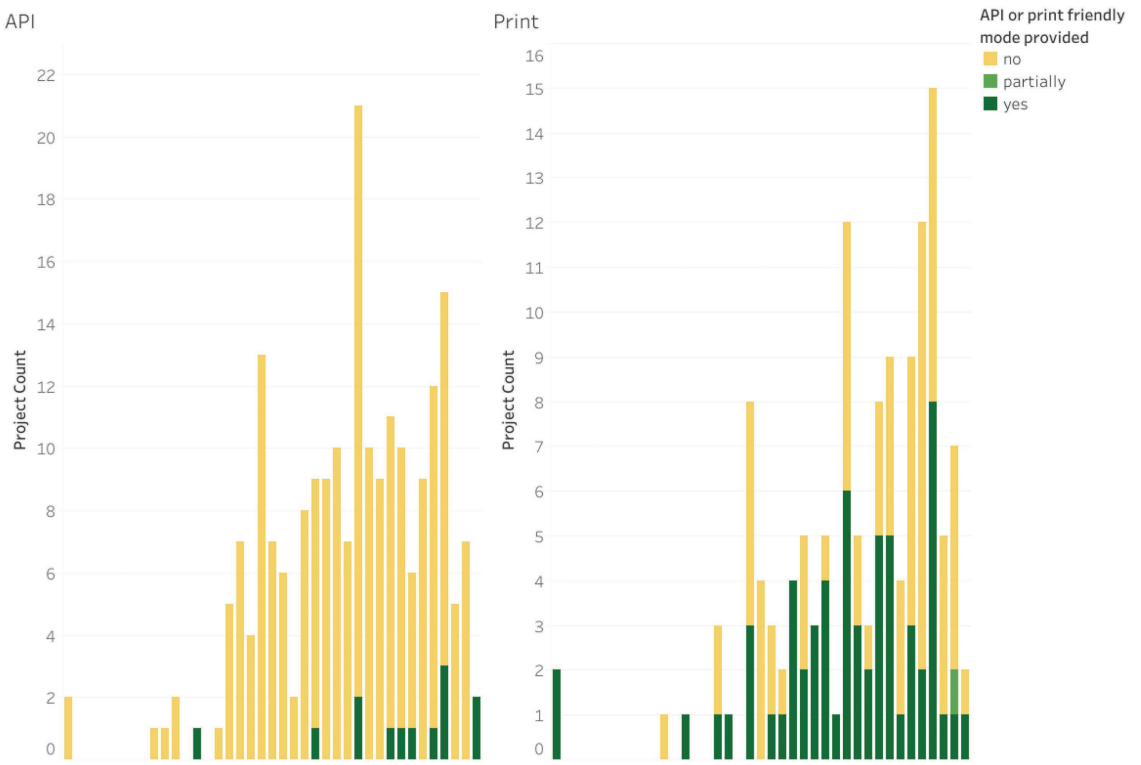

Figure 10. Availability of API or print function

Table 1: URL availability in Dig-Ed-Cat

May 2022 URL Status Check	Dig-Ed-Cat March 2022	% of Total Count
	None (Not available)	5.94%
	True (Available)	94.06%
404 (Not found)		6.79%
200 (Found)		93.21%

also due to ongoing loss.[11] To gain a better understanding of the longevity of digital editions, it was decided to undertake a more detailed analysis of all editions listed in the more extensive catalog maintained by Sahle. The goal was to produce an automated URL checker for each project URL listed by Sahle.

With the latest release 4.0, the underlying data for Sahle's catalog is openly available. The catalog data is a structured XML-TEI file, which means some work in Python is necessary to extract the desired data. The following section will outline the steps taken along with the code for this project.[12]

11. For a detailed discussion of loss among digital editions and efforts in preservation of DSEs, see the forthcoming white paper on the state of the field. Also see Brumfield (2018), Buddenbohm, Engelhardt, and Wuttke (2016), and Oltmanns et al. (2019).
12. The complete code is also available at https://github.com/mkrzmr/c21_catalogue.

```
from bs4 import BeautifulSoup as bs
import pandas as pd
!pip install lxml
```

This section installs the required libraries for this project. Beautiful Soup is a versatile parser for structured documents such as XML and HTML. While not a TEI-specific tool, in this case it can easily be used to find and extract data from TEI tags. Pandas is used to create DataFrames and perform basic operations on them. Beautiful Soup normally comes with a built-in XML parser, but in case the installation is without, the last line installs it.

A look at the XML file shows us what we are looking for.

```
<TEI xml:id="e1"><teiHeader><fileDesc><titleStmt><title>L'année 1437 dans
la pratique de Pierre Christofle, notaire du Châtelet d'Orléans - Digital Scholarly
Editions Catalog Entry</title><author>Catalog entry by PS</author></titleStmt>
<publicationStmt><publisher>Published by Patrick Sahle (Bergische Universität
Wuppertal) and the project team</publisher><availability><licence target="https://
creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0 (CC
BY 4.0)</licence></availability></publicationStmt><sourceDesc><p>Research data
collection, DSE catalog v. 4.0, based on earlier work (1996–2020)</p>
</sourceDesc></fileDesc><revisionDesc><listChange><change when="2018–
02–21" who="PS" type="creation"/></listChange></revisionDesc></teiHeader>
<text><body><bibl><title>L'année 1437 dans
la pratique de Pierre Christofle, notaire du Châtelet d'Orléans</title>
<rs type="sortkey">1437</rs><ref>http://elec.enc.sorbonne.fr/christofle/index.
html</ref><edition>Par Kouky Fianu avec la collaboration d'Anne Fortier.
Paris: École nationale des chartes, 2016.</edition><date type="firstPublication"
when="2016"/></bibl><p> "Pierre Christofle fut notaire
royal à Orléans de 1423 à 1450, attaché à la prévôté pour qui il rédigeait des
contrats portant le sceau de l'institution. Comme ses confrères, Pierre Christofle
inscrivait dans un registre et en une forme abrégée les conventions qu'il attestait.
Ces notes, au nombre de 387, rédigées entre le 1er janvier et le 29 décembre 1437,
conservées aux Archives départementales du Loiret sous la cote 3E 10144, sont ici
éditées dans leur ensemble. L'édition, indexée et prochainement téléchargeable,
permet l'enquête sur le lexique des actes." [from resource] </p><note type=
"labels">[hier stehen normalerweise Kommentare von mir zum internen Gebrauch]
</note><desc type="material" ana="charters">Kopiar, Notariatsurkunden</desc>
<desc type="subject" ana="history"/><desc type="language" ana="fr"/><desc type=
"era" ana="late_ma"/></body></text></TEI>
```

For each of the 789 entries, the data from the <title>, <ref>, and the <date> for the first publication should be extracted. The structure of XML-TEI makes it easy to extract the values from these tags. First, the document needs to be loaded and parsed—read—by Beautiful Soup to recognize the tags. At this point, we do not need to specify whether the document is encoded in TEI or not.

```
with open("catalog_TEI.xml", "r") as file: #load the file
bs_content = bs(file, 'lxml') #parse as XML
col ={'title':[],'url':[],'year':[]} #create columns for the DataFrame
df = pd.DataFrame(col) #create DataFrame
df.head() #check, should bring up empty DataFrame with columns now
```

The downloaded file "catalog_TEI.xml" is loaded and read by the XML parser. An empty DataFrame is created with the three columns *title*, *url*, and *year* as these are the values we are looking for at this point.

```
result = bs_content.find_all("ref") #simple search to export all URLs, even review links
for t in result:
print(t.text)
```

To show how well Beautiful Soup can process XML documents, this short statement is enough to find all <ref> tags and print out their text. Using t.text, we can select if we want the text of a tag or the attributes. This will become important later on. For now, the code prints out all links in the document (see Figure 11).

Immediately, some issues become apparent. Some links are links to reviews or further information about the edition, not the edition itself. Also, some plain

```
http://elec.enc.sorbonne.fr/christofle/index.html
http://www.abdn.ac.uk/bestiary
Catalog Entry Greta Franzini
http://apw.digitale-sammlungen.de/
http://synodes-protestants.symogih.org/index.html
http://www.cn-telma.fr/actesroyaux/index/
Catalog Entry Greta Franzini
http://www.masshist.org/digitaladams/archive/
https://rotunda.upress.virginia.edu/founders/ADMS.html
https://www.masshist.org/publications/jqadiaries/index.php
https://digital.janeaddams.ramapo.edu
http://www.philological.bham.ac.uk/Addison/
Catalog Entry Greta Franzini
http://www.forschungsdatenarchiv.escience.uni-tuebingen.de/adlils/adelbert_relatio
http://digital.nypl.org/schomburg/writers_aa19/
https://romantic-circles.org/editions/contemps/barbauld/poems1773/
Catalog Entry Greta Franzini

Catalog Entry Greta Franzini
```

Figure 11. Extracted links

text in the <ref> field was also extracted. This happened because we selected all <ref> tags, when only specific tags are of interest for this analysis. Making use of the hierarchical structure of XML, we can select only the <ref> tags within a <text> tag:

```
results = bs_content.find_all("text") #find all <text> tags
for entry in results:
output = [] #create empty list for output
title = entry.bibl.title.text #select the title by its tag
title = "".join(title.strip().split()) #Some titles have extra whitespaces, this removes them
url= entry.bibl.ref.text #again, select urls by the <ref> tag
date = (entry.find(type ="firstPublication")) #select only dates with the attribute "firstPublication"
year = (entry.bibl.date.get('when')) #select the value of the "when" field - the publication year
df.loc[len(df.index)]=(title,url,year)
```

Each <text> tag indicates a separate entry in the catalog. By searching for all and then iterating over the results, we can extract the required information for each entry. Beautiful Soup can navigate through the structure of the XML document and find the right <title> tag via its place in the hierarchy. As some titles extend over multiple lines and have line breaks and white spaces between them, the next line removes them. Similarly, URLs are extracted by only selecting the <ref> tag inside the <bibl> tag inside the <entry> tag. This removes reviews and URLs not related to this project.

Extracting the correct date is a bit more complicated: within the <date> tag, multiple attributes indicate first publication and relaunch, such as:
<date type="firstPublication" when="1996"/><date type="relaunch" when="2015"/>

Extracting the whole <date> tag would thus yield multiple dates. The solution is to first only select <date> tags that have the attribute "firstPublication" and then to extract the value from the "when" field. Finally, all three extracted values are added to the DataFrame (see Figure 12).

The next step is to prepare the URL status check. For this, we import a library that allows us to send out simple HTTP requests to the collected URLs.

```
import requests #this will allow us to check html status of the pages
df['status'] = "NaN" #add a new column to the dataframe
```

	title	url	year
0	L'année 1437 dans la pratique de Pierre Christ...	http://elec.enc.sorbonne.fr/christofle/index.html	2016
1	The Aberdeen Bestiary	http://www.abdn.ac.uk/bestiary	1996
2	apw - Acta Pacis Westphalicae	http://apw.digitale-sammlungen.de/	2014
3	Édition numérique des Actes des églises réform...	http://synodes-protestants.symogih.org/index.html	2018
4	Actes royaux	http://www.cn-telma.fr/actesroyaux/index/	2008
5	Adams Family Papers - An Electronic Archive	http://www.masshist.org/digitaladams/archive/	2003
6	The Adams Papers Digital Edition	https://rotunda.upress.virginia.edu/founders/A...	2008
7	John Quincy Adams Diary	https://www.masshist.org/publications/jqadiari...	2017
8	The Jane Addams Digital Edition	https://digital.janeaddams.ramapo.edu	2015
9	The Latin Prose and Poetry of Joseph Addison	http://www.philological.bham.ac.uk/Addison/	1997

Figure 12. Extracted DSE data

An empty column is added to the DataFrame; this is where the HTTP status report will go.

```
def check_url(url): #function to check all urls and report back
try:
request = requests.get(url, verify=False, timeout=10)
print ("Checked URL "+ str(url) +", Status was "+str(request.status_code))
return request.status_code
except requests.exceptions.RequestException as e: #basic error handling, if no response is
received, continue on
return e
```

Then, we need a basic function to check the HTML status of the pages. The above code tries to get a request status code. Because some pages will not produce one, we also need a way to deal with any connection errors we might encounter. A very simple solution is to just return the error message and continue on with the next URL.

```
df['status'] = df['url'].apply(check_url) #now checking all urls for their status code, might
take a minute
df.head() # checking one more time
df.to_csv('ouput.csv') #and output to a csv file for visualisation
```

With the above, the function is executed on each URL in the DataFrame, and the result is stored in the newly created *status* column. The result should look as shown in Figure 13.

	title	url	year	status
0	L'année 1437 dans la pratique de Pierre Christ...	http://elec.enc.sorbonne.fr/christofle/index.html	2016	200
1	The Aberdeen Bestiary	http://www.abdn.ac.uk/bestiary	1996	200
2	apw - Acta Pacis Westphalicae	http://apw.digitale-sammlungen.de/	2014	200
3	Édition numérique des Actes des églises réform...	http://synodes-protestants.symogih.org/index.html	2018	200
4	Actes royaux	http://www.cn-telma.fr/actesroyaux/index/	2008	404

Figure 13. DataFrame with completed URL check

With the last line, we export the DataFrame to a. csv file for any further data cleaning and processing. The three main steps in data cleaning in this example were the following:

1. In a few cases, no URL was in the catalog and so no check could be run. These few empty fields must be discarded for the analysis.
2. All connection errors are aggregated under the general *error* label. They count toward unreachable editions.[13]
3. It was found that a number of links referred to the Internet Archive. These links were marked up and visualized as a separate category.

Finally, the output can be plotted to show the percentage distribution of status codes for each year (see Figure 14). The results show that loss occurs for all years from 2021 onward (excluding pre-1994, where scarcity of sources skews the results somewhat). It confirms the findings from Dig-Ed-Cat but also shows how loss increases over time. If links referring to the Internet Archive are counted as signs of loss (which they should, for the Internet Archive's crawler has no access to the underlying database and is likely not able to capture an edition in its entirety), the picture is sobering. Pre-2004 editions experienced a loss rate of no less than 25%, with 50% of 1999 editions being unreachable.[14] In this brief case study, redirects are counted toward reachable sites, an assumption that is likely not correct in all cases as redirects will not always lead to the original resource. The lack of completely lost pages (404) in the pre-1996 years aligns with Sahle

13. For a detailed list of HTTP error codes, see the HTTP documentation at https://developer.mozilla.org/en-US/docs/Web/HTTP/Status. In the context of this study, all 200 status codes are successful connections, all 300 status codes are redirects, all 400 status codes are client-side errors, and all 500 status codes are server-side errors.
14. For this article, an edition is considered lost when it is no longer available under the known URL. Redirects to other sites are counted as available sites, although this might not always be the case. If both Dig-Ed-Cat and Sahle's list are seen as data sources for study and discovery, broken links in practice mean lost editions.

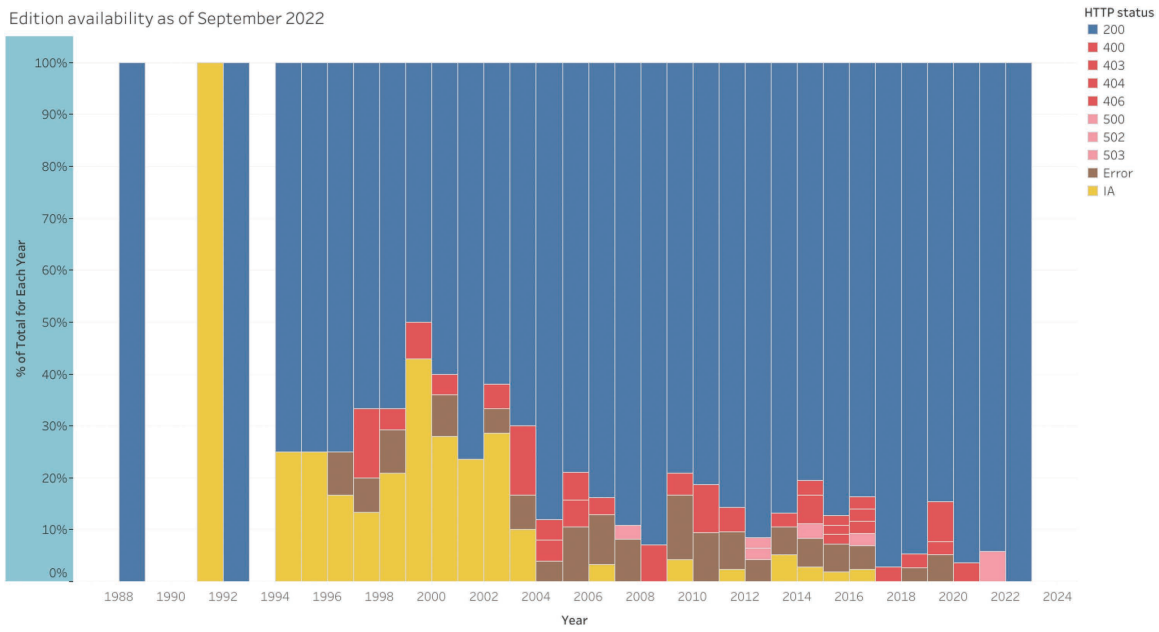

Figure 14. DSE availability, 1988 to September 2022

beginning work on the catalog in 1998. Earlier editions, already lost by then, probably never made it into the catalog.

A web crawler such as used by the Internet Archive's Wayback Machine captures the HTML content received from the server at the time of request. It generally does not capture the underlying XML data or any interactive elements such as witness selection or annotations. Depending on the crawler settings, it will follow a limited number of links on a website and may not capture every page in an edition.[15]

From a web archiving perspective, the fact that a large percentage of Internet Archive links are found shows that no widely used adequate preservation solution exists for digital editions. It is further unclear if the authors themselves deposited material into the Internet Archive or if, at a later point, curators used the Internet Archive to reconstruct a lost edition. In either case, these links must be seen as not providing the full edition in any way.

4. Conclusions

The purpose of this article was to demonstrate the application of a data-driven approach to the historiography of digital editions. It produced insight into the development of

15. For example, the Samuel Taylor Coleridge Archive is almost completely preserved in the Wayback Machine, while the Dread of the Rood Electronic Edition is partially preserved.

the field over time as well as provided a critique of the existing data sources. While both sources are very different in their approach (generally speaking, Sahle's list is a curated collection of editions, whereas the Dig-Ed-Cat is a crowd-sourced database), they were both very valuable sources of information for this study.

The data analysis was, where possible, focused on providing a longitudinal approach to assess the development of the field over time. Following a general overview of the number of projects in both data sources, results showed the average project duration decreased over time, while the number of projects per year kept increasing. An investigation of languages and project locations confirmed a strong focus on Europe and North America. Referring back to the documentation supplied with both datasets, this data bias was acknowledged by both sources. The data sources further have a strong bias toward open access academic editions, largely omitting proprietary editions.

As a consequence of this narrow focus in the available data, centralization of the field was reported as low. Investigating dominant access models, it could be shown that an uptrend in open access and CC licensing in the early 2000s predates changed funding requirements. Regarding encoding standards, the data analysis has shown that TEI became the dominant encoding model from 2012 on.

To understand the level of interoperability between digital editions, APIs in digital editions and export and print functionality were visualized as graphs. While APIs and print views are far apart in terms of functionality, they both represent functions for data export and potential data reuse in other contexts. The analysis was able to show that, while print functionality has been a part of digital editions throughout, APIs are only found in a few editions. These results indicate that interoperability between digital editions is low, with no widely used data exchange format and only a handful of APIs available.

As the final part, the analysis was concerned with the availability and sustainability of digital editions. Here, the larger number of editions in Sahle's list was chosen as the primary data source. The relevant data (edition name, year of first publication, and URL) were extracted from the structured data source. Necessary steps were described to enable reproducibility and increase transparency in the analysis. Distinguishing between plainly unavailable URLs, server errors, and links to the Internet Archive, loss rates could be determined to be around 5% per year, leading up to 25% to 50% of pre-2000 DSEs being unavailable. This result shows the need for a sustainable preservation solution for DSEs.

On a larger scale, the data sources drawn upon for this article show both the diversity of the field as well as the lack of comprehensive data sources on digital editions. While Dig-Ed-Cat is the largest source by number of records, it is reliant on community-provided entries (seven new editions were added between March 2022 and August 2023). The data fields used to describe editions are oriented toward textual

editions, a reasonable decision that at the same time complicates the recording of information on non-textual editions. As the analysis has shown, URLs change and resources are moved or deleted. To improve overall data quality, it might be prudent to periodically run a simple script such as the one described in this article to document the status of an edition's URL. Furthermore, where this indicates editions are lost and no copy can be obtained, a pragmatic solution would be to provide a link to an archived copy of the edition in a web archive. This would provide users with at least a snapshot of the original edition.

The results of this analysis contribute to a data-driven perspective on DSEs. They can inform debate about developments in the field, approximate what is lost already, and be meaningful impulses for the future development of digital editions. At the same time, it is a stark reminder that the problem of preserving DSEs is yet to be solved on a large scale. What editors may take away from this article is an increased awareness of longitudinal developments within digital editing as well as a data-based discussion of methods, licensing models, and interfaces. It is at the same time a challenge to editors to redefine the objectives of digital editing projects. Considering the longevity (or lack thereof) of digital editions, it may be called into question whether DSEs as a whole succeed in critically representing historical documents for scholarly engagement. In response to this challenge, editors might want to consider *availability* and *maintainability* of a DSE as important deliverables.

References

Andorfer, Peter, Ksenia Zaytseva, and Greta Franzini. "acdh-oeaw/dig_ed_cat: Release for Zenodo." 2018. https://doi.org/10.5281/ZENODO.1250797.

Boot, Peter, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini, and Dirk van Hulle, eds. 2017. *Advances in Digital Scholarly Editing: Papers Presented at the DiXiT Conferences in the Hague, Cologne, and Antwerp*. Leiden: Sidestone Press. https://www.sidestone.com/books/advances-in-digital-scholarly-editing.

Brumfield, Ben. 2018. "Preservable Digital Editions at AHA2018." FromThePage. https://content.fromthepage.com/preservable-digital-editions-at-aha2018/.

Buddenbohm, Stefan, Claudia Engelhardt, and Ulrike Wuttke. 2016. "Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum." *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2016_003.

Cummings, James. 2023. "Academics Retire and Servers Die: Adventures in the Hosting and Storage of Digital Humanities Projects." *Digital Humanities Quarterly* 17 (1).

Dalbello, Marija. 2011. "A Genealogy of Digital Humanities." *Journal of Documentation* 67 (3): 480–506. https://doi.org/10.1108/00220411111124550.

Dillen, Wout. 2019. "On Edited Archives and Archived Editions." *International Journal of Digital Humanities* 1 (2): 263–77. https://doi.org/10.1007/s42803-019-00018-4.

Driscoll, Matthew James, and Elena Pierazzo, eds. 2016. *Digital Scholarly Editing: Theories and Practices*. Cambridge: Open Book Publishers. https://doi.org/10.11647/OBP.0095.

Earhart, Amy E. 2012. "The Digital Edition and the Digital Humanities." *Textual Cultures* 7 (1): 18–28. https://doi.org/10.2979/textcult.7.1.18.

European Commission. 2021. "EU Grants: AGA – Annotated Grant Agreement: EU Funding Programmes 2021–2027." https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf.

Franzini, Greta. (2015) 2022. "How to Contribute a Digital Edition to the *Catalogue*." digEds_cat. https://github.com/gfranzini/digEds_cat/blob/81df723e147e50cb68fc1eb4393b6b56cd8209e7/CONTRIBUTING.md.

Franzini, Greta, Melissa Terras, and Simon Mahony. 2019. "Digital Editions of Text: Surveying User Requirements in the Digital Humanities." *Journal on Computing and Cultural Heritage* 12 (1): 1–23. https://doi.org/10.1145/3230671.

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine, with Rachel Donahue. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. CLIR Publication, no. 149. Washington, DC: Council on Library and Information Resources.

Oltmanns, Elias, Tim Hasler, Wolfgang Peters-Kottig, and Heinz-Günter Kuper. 2019. "Different Preservation Levels: The Case of Scholarly Digital Editions." *Data Science Journal* 18 (October). https://doi.org/10.5334/dsj-2019-051.

Risam, Roopika. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston, IL: Northwestern University Press.

Risam, Roopika, and Alex Gil. 2022. "Introduction: The Questions of Minimal Computing." *Digital Humanities Quarterly*. https://www.semanticscholar.org/paper/Introduction%3A-The-Questions-of-Minimal-Computing-Risam-Gil/d6facb14aefbe04cac859a96817b705326d821bb.

Sahle, Patrick. 2016. "What Is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19–40. Cambridge: Open Book Publishers. https://doi.org/10.11647/OBP.0095.02.

———. 2020. "About." A Catalog of Digital Scholarly Editions v 4.0. https://digitale-edition.de/exist/apps/editions-browser/about.html.

Siddiqui, Nabeel. 2022. "Hidden in Plain-TeX: Investigating Minimal Computing Workflows." *Digital Humanities Quarterly*. https://www.semanticscholar.org/paper/Hidden-in-Plain-TeX%3A-Investigating-Minimal-Siddiqui/3ff44301b804a63d7548a4c6b2ab16745f58a453.

Unsworth, John. 1998. "Documenting the Reinvention of Text: The Importance of Imperfection, Doubt, and Failure." https://web.mit.edu/m-I-t/articles/unsworth.html.

Zundert, Joris J. van. 2016. "Close Reading and Slow Programming—Computer Code as Digital Scholarly Edition." In *ESTS 2016 / DiXiT 3*. Letterkunde (HI). https://pure.knaw.nl/portal/en/publications/ee62365a-b6e3-4ce4-8511-9f1feab97003.