Some Digital Editions and Some Remaining Challenges

Peter Boot

Huygens Institute. Royal Netherlands Academy of Arts and Sciences.

Resumen

La edición digital ha logrado enormes progresos desde la llegada de Internet; sin embargo, hay muchas cosas que todavía no sabemos cómo conseguir. Examino en este trabajo una serie de ediciones digitales llevadas a cabo y lo que nos han enseñado. A continuación, trato de algunas cuestiones planteadas en relación con las ediciones digitales: cómo involucrar a voluntarios en la creación y el mantenimiento de la edición, la forma de reducir el costo de la creación de las ediciones y cómo lidiar con las amenazas a su conservación en el tiempo.

Abstract

While the digital edition has made enormous progress since the advent of the web, there is also much that we still don't know how to achieve. I examine a number of editions and some things they have taught us. Then I discuss a few open questions regarding digital editions: how to involve volunteers in the creation and upkeep of the edition, how to lower the cost of creating editions and how to deal with threats to their longevity.

Palabras clave: Ediciones digitales, Longevidad digital, Normalización, Tercerización masiva

Keywords: Digital eitions, Digital longevity, Standards, Crowd sourcing

INTRODUCTION

In the period since the advent of the World Wide Web, we have learned much about the creation of online editions and online collections of historical literature. We have learned how to create enduring digital representations of historic books and learned how these representations can

be shown online. We learned how to create navigational facilities such as thumbnails and tables of contents. We learned how to add images and zooming facilities to digital editions. We learned how to index them and how to make them searchable. We are learning how to engage larger communities in the creation of digital resources. Emblem scholars have been particularly active in the process, witness e.g. the early emblem databases developed in A Coruña by the group of Sagrario López Poza.

There are also, however, many things that as yet we do not know about digital editions. There are still unsolved issues with respect to our editions' longevity. Digital editions are far too expensive to produce and maintain. A wider question is how we can turn static editions into community resources that can be extended and annotated by scholars, students, and other interested persons.

In this article, I will discuss a number of digital editions that I have been involved with, and mention some of their stronger points and some of their limitations. I will then examine some of the wider issues digital edition developers are facing, and point at some questions that merit further investigation.

EMBLEM PROJECT UTRECHT

The Emblem Project Utrecht (EPU) was created by Els Stronks at Utrecht University. The EPU digitized 27 books of Dutch love emblems. The love emblem was an emblematic subgenre that originated in the Netherlands in the first years of the seventeenth century and enjoyed a wide success in other European countries (Sebastián López, 2001). Among emblem projects, the EPU was ground-breaking in at least three respects: the project was the first to create full transcriptions of a collection of emblem books and to encode these in TEI/XML, it made available high-resolution images of all pages, and it decided to index emblem pictures with Iconclass. See Fig. 1 for a typical page from the project. Shown is an emblem from P.C. Hooft's *Emblemata amatoria* (Amsterdam, 1611), with *pictura*, transcription with links to translations and on the left the table of contents, including thumbnails. The table of contents includes links to other components of the edition and the emblem representation.[1]

---

[1] About the project: (Daly, 2007; Stronks and Boot, 2007). I was involved in the earlier stages of the project and responsible mostly for the technical set-up.

The decision to transcribe the full text using TEI/XML was important for a number of reasons (Boot, 2004). Earlier emblem projects had mostly limited themselves to transcribing emblem mottoes, which severely limits searchability and prejudges the issue of the relative importance of the emblem's various constituents. That the transcription should use a structured format such as XML was especially important for a genre as full of internal structure as the emblem book. The use of Iconclass for indexing the content of the pictures was important among other reasons because Iconclass provides a hierarchical system where indexers don't have to choose between the more generic term ('bird') and the more specific one ('eagle'), because its notations are to a large extent language-independent, and because it contains ready-made categories for most of the pictorial motifs that were used in emblem books (Brandhorst, 2004). Many of the choices in emblem digitization made by EPU were followed by later emblem projects, such as the *French emblems at Glasgow* site.[2] (This is not to say we didn't also make some less fortunate decisions. For instance, the XML encodings we devised were in some respects much too complicated, leading to mistakes and lessened productivity) (Fig. 1).

It is interesting to note that some of the site's limitations (or what can be seen as such) follow from the decision to specialize. The site's layout and navigation make sense for emblem books. As long as the researcher's interest is focused on the emblem and its relation to other emblem books, this presents no problem. If the interest moves towards the wider culture of which the emblem is a part (songbooks, non-illustrated literature, manuscript additions), the site's design may become a straitjacket. This is not a fault of the site designers (us), but a case of the general law that it is hard to build sites that are generally applicable and useful. If it is the specific properties of a genre that help organize and navigate a site, then the site loses focus and usefulness if the properties no longer obtain. We will see the same issue return below.

GROTIUS CORRESPONDENCE

Unlike the EPU, which was created from scratch in a digital environment, Grotius' correspondence had first been edited in a series of books. Hugo Grotius, a seventeenth-century poet, statesman, and lawyer, best known for his seminal contributions to international law, had a large network of correspondents all over Europe. His correspondence was edited in 17 volumes in a project that took more than seventy years to complete

---

[2] http://www.emblems.arts.gla.ac.uk/french/.

(1928-2001). Huygens ING decided to digitize the books and the results were published online in 2009.

Basic digitization was done by the Digital Library of Dutch Literature (DBNL). DBNL has a well-tested production line where thousands of Dutch titles have been efficiently processed for publication at http://www.dbnl.org/. DBNL creates a light XML encoding geared towards a web publication that more or less re-creates the original book form. The Grotius volumes are available on the DBNL site in that form (see e.g. http://www.dbnl.org/tekst/groo001brie01_01/). The site gives a representation of the original volumes with an electronic table of contents. While this is certainly in many respects an improvement over the physical books (if only because they can be searched and can be consulted from anywhere in the world), it is also a very limited representation, as there is in this edition no concept of letters, correspondents or dates. A digital edition of correspondence should, at the very least, facilitate selection by year and correspondent. The edition should not be organized in terms of the book volumes in which it was originally published.

In the representation of the correspondence that was developed at Huygens ING we did organize the correspondence as a database of letters, where users can indeed make selections by year, sender or recipient. Fig. **2** shows the letters that Grotius (Swedish ambassador in Paris) sent to Queen Christina of Sweden.

What we did not decide to do, however, is to use the existing material and create a new digital edition that would hide all traces of ever having been a book. The user, if he wishes to do so, can still select letters by the volume in which they were published (fig. 3). The seventeen introductions and bibliographies of the original volumes are also available. More embarrassingly, there are also seventeen indices that have not yet been hyperlinked to the pages that they point to. Neither did we apply any corrections.

So what we have here is an edition that very visibly straddles the old and the new, even in language: the new navigation is targeted towards an international audience and uses English, while the old annotation is still in Dutch. The resulting hybrid edition could not without considerable expenditure be reworked into a modern, fully integrated edition, even though it would be very much desirable to do something about some of its more obvious shortcomings.

LETTERS OF VINCENT VAN GOGH

The correspondence of Vincent Van Gogh was published jointly by the Van Gogh Museum and Huygens ING in 2009. The edition differs from both EPU (planned as digital edition) and Grotius (original publication in book form) in that it was simultaneously published on both platforms. During the preparation of what was supposed to become a series of books, the web developed into the most suitable publication medium for a scholarly edition of the letters. As it was recognized that many people would still want to read the letters in book form, the decision was made to create a dual publication: both a full scholarly edition, freely available online, and a lavishly illustrated six-volume book edition where some of the scholarly detail is omitted. The book edition was published in three languages: Dutch and French (the original languages of the letters) and English. All of these versions were published simultaneously. For more background see (Boot, 2011b).

A prototypical edition page is shown in figure 4. The letters are shown in a two-column layout. The reader can choose what he wants to see in the columns: the original text, a facsimile, a translation, notes or thumbnails of the works of art discussed in the letters. The menu is visible at the top of each page and gives immediate access to tables of contents, search functionality and explanatory or expository material.

Most digital edition projects are probably built from the outset in a structured digital environment. The Van Gogh letters were initially edited in a word processor environment (MS Word), creating the need for a conversion into XML. The conversion was a time-consuming process, because of the many different ways a similar visual effect can be accomplished in MS Word. Many manual checks and corrections were needed before the result was acceptable.

Still, the conversion was in many ways an instructive process, among other things because it taught us that many of the encoding aspects that we tend to take for granted may not be really necessary. If one sets out to create a digital edition of correspondence in TEI, one will probably mark-up datelines and salutations and address lines. Our conversion could not distinguish plain paragraph text from salutations and other special content, and therefore encoded all text as 'anonymous blocks'. This is certainly not the sort of encoding that one would expect in a somewhat high-profile TEI project, but it has not caused any problems.

Another important lesson from the Van Gogh project is that a text's full online availability does not necessarily hurt book sales. Confirming the evidence of other publishers (Doctorow, 2006; Hilton III and Wiley, 2010),

the sales of the book volumes, which were anything but cheap, was quite satisfactory. Even though log analysis shows that there are people who read the online edition sequentially (Boot, 2011a), most people probably prefer the book for sustained reading.

REMAINING CHALLENGES

A characteristic that all three discussed projects share –or a limitation if you will– is that they all were created as finished products. A small group of people designed and created the site, the rest of the world can either like it or not, but does not have the option of improving the site, either through addition of primary material, through correction or through annotation. In that sense, these websites are like books, which after their creation move to library shelves to be consulted by readers.

An electronic book, available to everyone, is not of course a bad thing, but the promises of hypertext extended much beyond a collection of read-only, discrete entities. An important remaining challenge for the field of digital editing is how to overcome the static character of present-day digital editions and turn them into the dynamic repositories of textual knowledge as sketched by Peter Robinson (2005) and, under the label of the 'knowledge site', by Peter Shillingsburg (2006). Robinson has even argued that the adoption of a digital edition by a community of scholars is the best guarantee of its long-term survival (2003). Further discussion of the role of volunteer contributors and the social edition is given in (Gibbs, 2011) and (Timney, Leitch, and Siemens, 2011).

Up to now, most attempts to involve a larger group of people in in editorial work have centred on getting non-experts to participate in transcription tasks. At Huygens ING, a group of trained volunteers transcribed a late medieval encyclopaedia (Versélewel de Witt Hamer, 2009). Elsewhere in the Netherlands, a group of volunteers transcribed a number of early Bible translations. The best-known of this class of projects is Transcribe Bentham, the initiative based at University College London to get the public to participate in the transcription of Jeremy Bentham's papers. As appears from the project's recent report, an appeal to the crowd is certainly not by itself a simple and free answer for large volumes of transcription work (Causer, Tonra, and Wallace, 2012). One of the issues stressed by Causer et al. is that volunteers should receive feedback and should get the impression that their contribution is being valued. Either project staff should be available for that purpose (thus spending the financial resources that employing volunteers was meant to save), or the task of providing guidance and feedback to volunteers should itself be outsourced to

volunteers. Another problem faced by Transcribe Bentham was presented by the XML encoding that participants were expected to deliver (necessary for representing such things as manuscript additions and deletions). Even though the project provided a toolbar for creating the required XML tags, many volunteers apparently found the encoding complicated and unappealing. We will come back to the issue of XML encoding below. As Causer et al. mention, one way of dealing with the issue might have been to provide a visual representation of the xml-tags: rather than <del>-tags (for deletions), users would have seen crossed-out words.

These difficulties apart, the employment of volunteers in the preparation of what is otherwise a traditional and static edition is still a far cry from the edition run and maintained by volunteer contributors as proposed by Robinson. There are at least two obstacles to wider involvement of volunteers in the maintenance of the digital edition: editorial culture still resists the idea of amateur participation, and the digital edition's production process is centralized and not designed for public participation. It is true that, if outsiders would feel welcome and would have the tools to participate, there might still be only a limited number of people with a sufficient interest in the text. Vanhoutte has argued that for most textual traditions, especially in smaller language areas, the community model is improbable, simply because the communities would be too small (2012). That is undoubtedly true, but it is also true that high barriers to participation will discourage people who might otherwise be interested. I will not discuss the aspects of editorial culture which resist outsider participation. It may be that editors tend to be on the defensive because people often don't understand that editing is a profession that requires expertise and is not something that anyone could do. I will examine, however, the role of the edition's production process.

For most of the success stories of mass online collaboration (message boards, weblogs, Wikipedia, Flickr), a decisive factor in their success was that a web browser was all that was necessary for participation. If uploading a photo would have required using an FTP client, Flickr would not have enjoyed a tenth of its success. The current way of preparing a digital edition, on the contrary, is very much an off-line process and requires unfamiliar tools and uncommon technical expertise. Typically, it involves the creation of XML files and the development of display programs specific to a particular edition and dependent on its specific encoding guidelines. The XML files tend to be locked away from the public. And while XML provides superior technology for the creation of digital editions, there is no denying that many people would rather have nothing to do with it. The resistance is based on theoretical grounds, as well as on practical issues:

XML is seen as complex, time-consuming and distracting from the intellectual questions that motivate editorial work. What we should try to work towards, perhaps, is a combination between XML content and a production process for digital editions along the model of the CMSs (Content Management Systems) that power most of today's web sites: they require no special knowledge of the language of the web (HTML), they are managed from within a web browser and require no other locally installed software, and content that is entered only once can be reused in different formats.

In fact, over the last few years big strides have been made towards the integration of the XML and CMS worlds for the digital edition. The eLaborate toolset developed at Huygens ING (Beaulieu, Dalen-Oskam, and Zundert, 2013) is a CMS-like transcription and annotation environment, the content of which has been exported to TEI/XML. Most current efforts in this direction seem to be based on extending the Drupal CMS for working with TEI documents.[3] The most technically advanced of these initiatives is at present probably TEICHI, the environment initially developed for the publication of Bérardier de Bataut's *Essai sur le récit* (Pape, Schöch, and Wegner, 2012). One of the promising aspects of the TEICHI approach is that different parts of a work can be uploaded as separate XML files, which creates the possibility of multiple people working concurrently on an edition of the same work. An online TEI editor is planned for the future.

Closely related to the issue of volunteer participation are the issues of cost and of long-term maintenance. On the question of cost, for all three editions discussed above it took serious money to develop them. The Van Gogh edition was the most expensive, among other things because it was the most heavily researched, but even without any research[4] planning, transcription, programming, indexing, design and testing cost money. EPU could be built thanks to a grant from the Dutch Organisation for Scientific Research (NWO), costs for the digitization of the Grotius letters were borne by Huygens ING, and the Van Gogh site was paid for by the Van Gogh Museum and Huygens ING, supported by an array of sponsors. One of the comments that has been made with regards to the publication of the Van Gogh letters is that it "[p]aradoxically (…) could have a negative impact on digital scholarly editing in the future: (…)pushing the standards so far that

---

[3] As evidenced by discussions on the TEI and Drupal mailing lists, the Canadian IslandLives project (http://www.islandlives.ca/) uses TEI in a Drupal setting. The Tapas project (http://www.tapasproject.org/) has also announced it will use Drupal in its projected TEI hosting and publication services.

[4] The research for the Grotius edition was obviously also expensive but had already been done for the earlier publication on paper.

only a happy few, owing to the financial implications, would be able to follow suit" (Van Raemdonck, 2012). This is a valid complaint. It would be a sorry state of affairs if excellent scholarly editing should be reserved for cultural icons of Van Gogh's level. The only way of avoiding that situation is to try to lower the costs of digital editing.

For both the EPU and the Van Gogh site, the user interface was custom developed for these specific sites. It is clear that for any edition, the nature of the available material will be to some extent unique, and will require a specific presentation. What we should try to achieve is that the larger part of a typical edition's functionality can be assembled from standard building blocks, and only the specific five or ten percent of the functionality needs to be developed from scratch. With the increasing number of high quality digital editions showcasing different types of texts and different sorts of editorial approaches, it should be possible to abstract a limited number of models for the digital edition, and develop software capable of handling a generic edition of, say, letters, or emblem books.

Ideally, this generic software would be developed in cooperation between multiple editorial institutes. If that should be impossible, we should at least agree on a set of rules that can guarantee interoperability between the components developed at the various editorial institutes. This is in fact one of the aims of the InterEdition project, that was funded as an EU COST action from 2008 to 2012. An recurring problem is that effective interoperability requires agreement about standards, which are time-consuming to establish and which require discipline to maintain. The aim should be to reach a situation where adhering to standards, rather than being one more thing to worry about, could actually make developing the edition easier and cheaper, not just in the publication stage but from the earliest phases of development.

We need these shared publication platforms not just to bring down the initial development costs. They are quite as important from a maintenance perspective. For an institute such as Huygens ING, which aims to publish multiple digital editions each year, the maintenance burden would quickly outgrow the development budget if each digital edition would have its own publication platform and would therefore need separate maintenance.

That brings us to the issue of the long-term availability of our digital editions. The fact is that even though XML has mostly, but by no means everywhere, been accepted as the natural language for encoding digital editions, and even though TEI has, to some extent, been successful in creating uniform encoding practices, the availability of digital editions for the longer term is by no means assured. Every day, servers are being decommissioned, domain name registrations lapse, server software is

upgraded, new versions of web browsers are released, programmers leave departments or departments are reshuffled altogether or even disappear. All of these events constitute threats for the continued availability of digital editions. While there is no simple response to these threats, one important ingredient in any solution is to keep things as simple as possible: to reduce the number of software and hardware components that an edition depends on, to reduce the amount of expertise required to keep the edition running, sometimes to resist the temptation of the newest and shiniest and stick to the stable and well-tested.

While we have no way of knowing what the computer and the web will look like in, say, fifty years, one thing that we know is that it will be radically different from what we know today. Thirty years ago, the personal computer was in its infancy, twenty years ago there was no World Wide Web, ten years ago mobile Internet access was a rarity. It is hard to prepare for unknown circumstances. When we are asking for shared publication platforms, for the application of standards and for a self-imposed limitation to use proven technology only, we may be at odds with the competitive nature of the scientific process and ultimately human nature itself. Both researchers and institutions want to distinguish themselves by showing what they can do, rather than just do what others have done or are doing. A better answer to the volatility of the technological and organizational environment might be to create low-technology backup solutions for the day when today's shiny interfaces no longer work. That low-technology alternative might even be a book, printed in a limited number of copies. Or, less radical, depositing the edited texts with a general purpose digital library, assuring the long-term availability of at least the edition's intellectual content.

CONCLUSION

While the development of the personal computer and the web has brought many innovations to the digital edition, there is still ample room for improvements. As we have seen, the creation and maintenance of digital editions is still a job for specialists and the underlying technology does not really support external participation. The prohibitive costs of creating digital editions are another factor that is keeping out potential participants. While research itself will always remain expensive, it should be possible to create production lines for building digital editions that are more welcoming to occasional or non-academic participants. Standards are an essential ingredient in such production lines. However, as standards are always the embodiment of past research, standards by themselves can't guarantee our editions' long-term survival, and we should also think creatively about fall-

back solutions for the preservation of editorial work in a mostly paperless environment.

Literature

Beaulieu, Anne, Karina van Dalen-Oskam, and Joris van Zundert, "Between tradition and Web 2.0: eLaborate as social experiment in humanities scholarship" en *Social Software and the Evolution of User Expertise: Future Trends in Knowledge Creation and Dissemination*, Tatjana Takševa (ed.), IGI Global, 2013.

Boot, Peter, "Accessing Emblems using XML. Digitisation Practice at the Emblem Project Utrecht" en *Florilegio de Estudios de Emblematica. A Florilegium of Studies on Emblematics. Actas del VI Congreso Internacional de The Society for Emblem Studies. Proceedings of the 6th International Conference of the Society for Emblem Studies. A Coruña, 2002*, Sagrario López Poza (ed.), Ferrol, Sociedad de Cultura Valle Inclán, 2004, pp. 191-197.

Boot, Peter, *Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship*, Utrecht, Utrecht University, 2009.

Boot, Peter, "Reading Van Gogh Online?" en *Ariadne* 66 (2011a), <http://www.ariadne.ac.uk/issue66/boot/>, [2012-05-20].

Boot, Peter, "Vincent Van Gogh - The Letters" *Jahrbuch für Computerphilologie - online* (2011b), <http://computerphilologie.tu-darmstadt.de/jg09/boot.html>, [2012-05-20].

Brandhorst, Hans, "Using Iconclass for the Iconographic Indexing of Emblems" en *Digital Collections and the Management of Knowledge. Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images*, Mara Wade (ed.), DigiCULT, 2004, pp. 29-44.

Causer, Tim, Justin Tonra, and Valerie Wallace, "Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham" en *Literary and Linguistic Computing* 27 2 (2012). pp. 119-137.

Daly, Peter M., "Digitising Dutch Love Emblems" en *Learned Love*, Els Stronks and Peter Boot (eds.), The Hague, DANS, 2007, pp. 183-200.

Doctorow, Cory. 2006. Giving It Away. *Forbes* 1-12-2006.

Gibbs, Frederick W., "New Textual Traditions from Community Transcription" en *Digital Medievalist* 7 (2011).

Hilton III, John , and David Wiley, "The Short-Term Influence of Free Digital Versions of Books on Print Sales" *Journal of Electronic Publishing* 13 (1) (2010).

Pape, S., C. Schöch, and L. Wegner, "TEICHI and the Tools Paradox. Developing a Publishing Framework for Digital Editions" en *Journal of the Text Encoding Initiative* (2) (2012).

Robinson, Peter, "Where we are with electronic scholarly editions, and where we want to be" *Jahrbuch für Computerphilologie* 5 (2003), pp. 125-146.

Robinson, Peter, "Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?" en *Digital Medievalist* 1 (1) (2005).

Sebastián López, Santiago, *La mejor emblemática amorosa del barroco*, *Collección Sielae*, Sociedad de Cultura Valle Inclán, 2001.

Shillingsburg, Peter L., *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge, Cambridge University Press, 2006.

Stronks, Els, and Peter Boot, "The Dutch love emblem on the Internet: an introduction" en *Learned Love*, Els Stronks and Peter Boot (eds.), The Hague, DANS, 2007, pp. 1-9.

Timney, Meagan, Cara Leitch, and Ray Siemens, *Opening the Gates: A New Model for Edition Production in a Time of Collaboration*, <http://etcl.uvic.ca/files/2011/01/timneyleitchsiemens-socialedition.pdf>, [2011-04-10].

Van Raemdonck, Bert, "[Review of] Vincent van Gogh - The Letters" *Variants* 8 (2012), p. 227-230.

Vanhoutte, Edward, *Being Practical. Electronic editions of Flemish literary texts in an international perspective*, <http://edwardvanhoutte.blogspot.com/>, [2012-05-20].

Versélewel de Witt Hamer, Noor, *Bartholomeus Engelsman, Van den proprieteyten der dinghen. Een diplomatische editie van de Middelnederlandse vertaling (1485) van de 13de-eeuwse encyclopedie De proprietatibus rerum van Bartholomaeus Anglicus*, <http://bartholomeusengelsman.huygens.knaw.nl/path>, [2012-05-20].
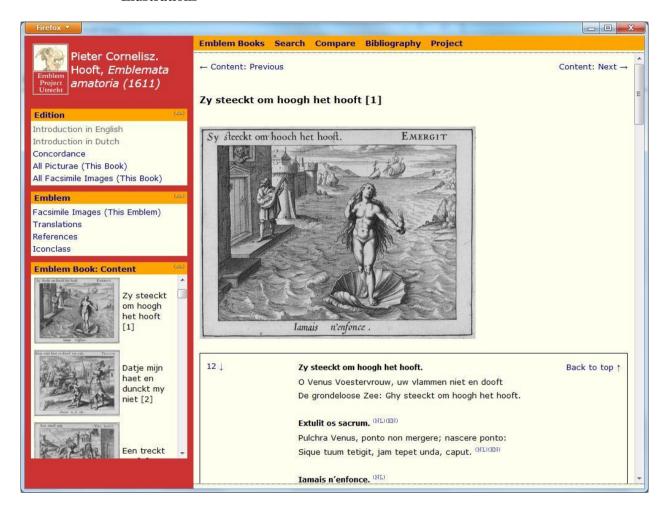
Illustrations

Fig. 1: Emblem Project Utrecht site, showing an emblem from P.C. Hooft's *Emblemata amatoria*
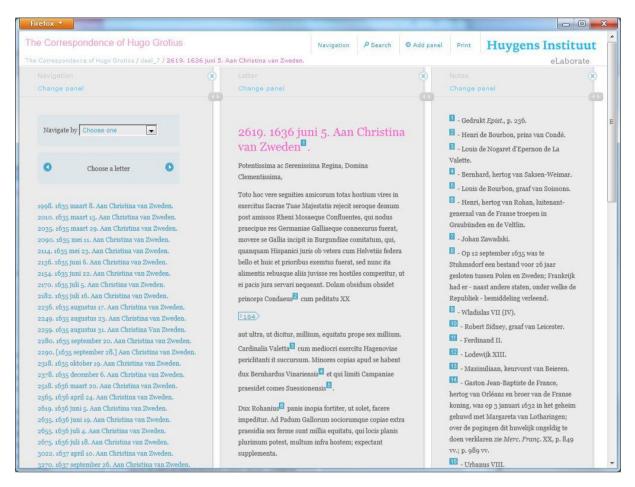
Fig. 2: Correspondence of Hugo Grotius. Shown is a letter to Queen Christina of Sweden, with on the left a list of all letters that Grotius wrote to her.
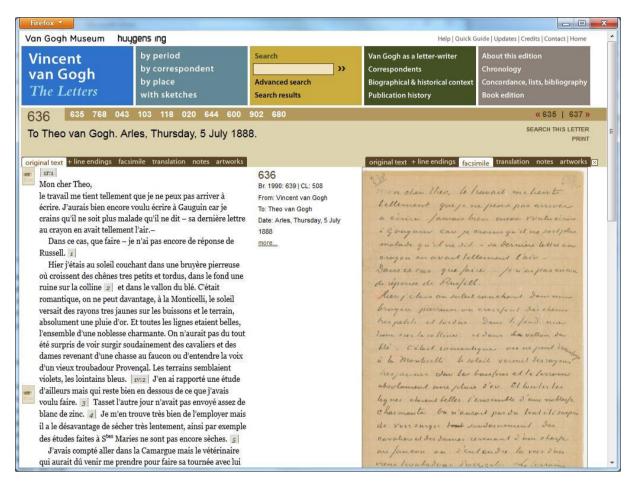
Fig. 3: Navigation options in the Grotius edition include volumes.

Fig. 4: Van Gogh edition.