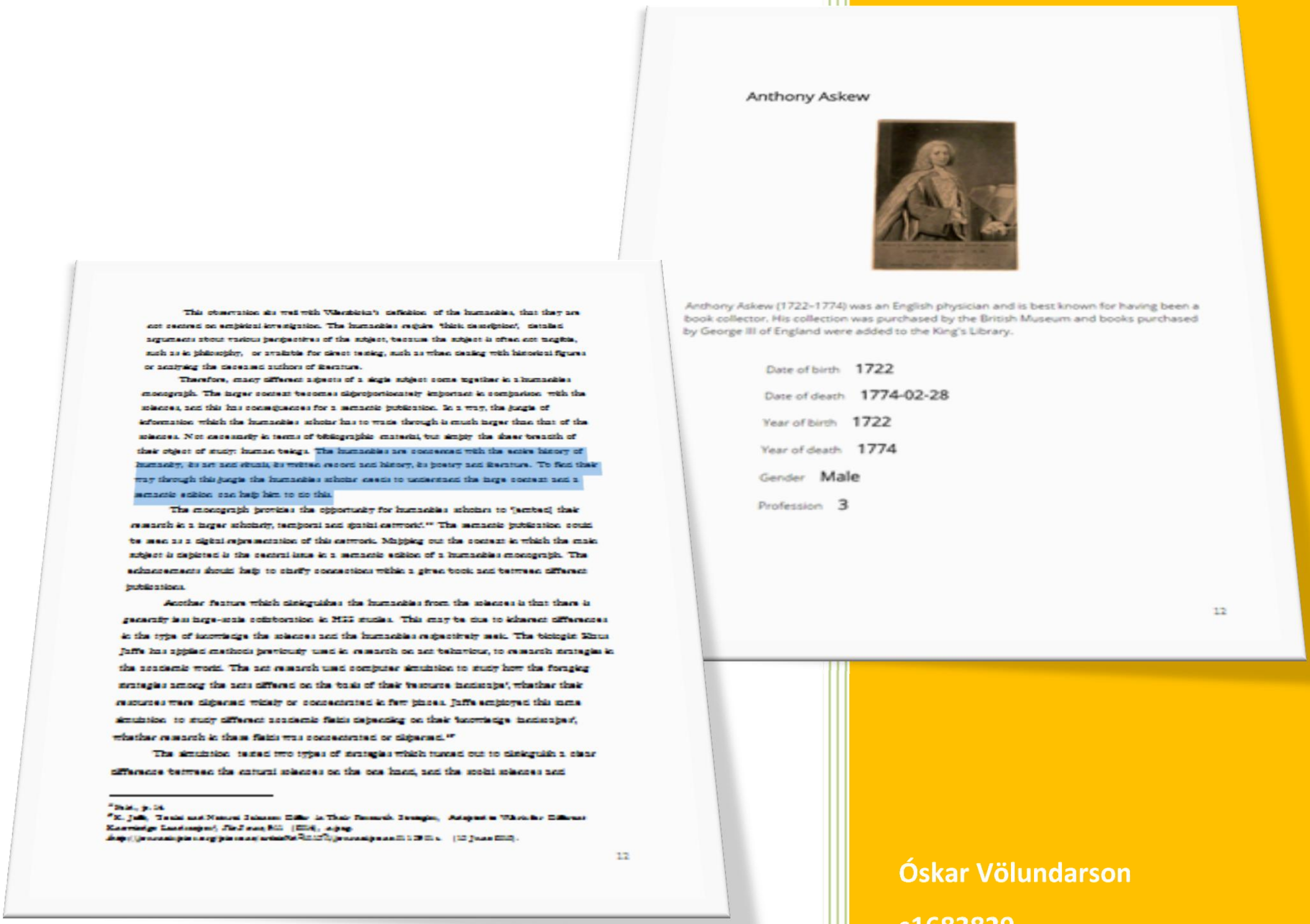
Semantic Publishing in the Humanities
Enhancing the reader's experience



This observation fits well with Wierzbicka's definition of the humanities, that they are not centred on empirical investigation. The humanities require 'thick description', detailed arguments about various perspectives of the subject, because the subject is often not tangible, such as in philosophy, or available for direct testing, such as when dealing with historical figures or analysing the deceased authors of literature.

Therefore, many different aspects of a single subject come together in a humanities monograph. The larger context becomes disproportionately important in comparison with the sciences, and this has consequences for a semantic publication. In a way, the jungle of information which the humanities scholar has to wade through is much larger than that of the sciences. Not necessarily in terms of bibliographic material, but simply the sheer breadth of their object of study: human beings. The humanities are concerned with the entire history of humanity, its art and rituals, its written record and history, its poetry and literature. To find their way through this jungle the humanities scholar needs to understand the large context and a semantic edition can help him to do this.

The monograph provides the opportunity for humanities scholars to '[embed] their research in a larger scholarly, temporal and spatial network.'[illegible] The semantic publication could be seen as a digital representation of this network. Mapping out the context in which the main subject is depicted is the central issue in a semantic edition of a humanities monograph. The enhancements should help to clarify connections within a given book and between different publications.

Another feature which distinguishes the humanities from the sciences is that there is generally less large-scale collaboration in HSS studies. This may be due to inherent differences in the type of knowledge the sciences and the humanities respectively seek. The biologist Klaus Jaffe has applied methods previously used in research on ant behaviour, to research strategies in the academic world. The ant research used computer simulation to study how the foraging strategies among the ants differed on the basis of their 'resource landscape', whether their resources were dispersed widely or concentrated in few places. Jaffe employed this same simulation to study different academic fields depending on their 'knowledge landscapes', whether research in these fields was concentrated or dispersed.[illegible]

The simulation tested two types of strategies which turned out to distinguish a clear difference between the natural sciences on the one hand, and the social sciences and

[illegible] Ibid., p. 14.

[illegible] K. Jaffe, 'Social and Natural Sciences Differ in Their Research Strategies, Adapted to Work for Different Knowledge Landscapes', PloS one, [illegible] (2014), [illegible] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.[illegible] (12 June 2016).

12

Anthony Askew

Anthony Askew (1722-1774) was an English physician and is best known for having been a book collector. His collection was purchased by the British Museum and books purchased by George III of England were added to the King's Library.

Date of birth	1722
Date of death	1774-02-28
Year of birth	1722
Year of death	1774
Gender	Male
Profession	3

12

Óskar Völundarson

s1683829

Book and Digital Media Studies

Supervisor: Peter Verhaar

Second reader: Adriaan van der Weel

Date of completion: 18 July 2016

Word count: 20.101

Table of Contents

Abstract

Digital technology enables us to access and examine texts in ways that are not possible in printed publications. One of the potential digital enhancements involves making the meaning of texts machine-readable. This has been referred to as *semantic publishing* and many scientific publishers have made extensive use of semantic technologies in their publications. Meanwhile, the potential of semantic enhancements for the humanities remains to a great degree unexplored. This thesis examines semantic enhancements in the context of how humanities research is conducted: Which type of humanities publication is best suited for semantic enhancement? Which guidelines should govern how the text is coded? And how can the end-users of the book benefit from the enhancements? These questions are examined through a case study of a single monograph (*The Book-hunter in London*, 1895) since this is a particularly important form of publication in the humanities. The focus throughout is on the end-user of the enhanced edition.

Introduction

Books are carriers of information.[1] Even publications that are chiefly meant for entertainment or aesthetic pleasure, such as crime fiction and coffee table books, are essentially collections of data. They relate information on a specific topic to the reader through text or images. Some genres of books are more directly concerned with relating information than others. Among them is the academic book. Its primary purpose is to inform the reader and it should be designed to make this information retrieval as easy and efficient as possible.

The modern printed book in many ways lives up to this task. Since the Middle-Ages, bookmakers have been developing various user-friendly features[2] such as indexes, underlining, and page numbers. Useful as these features are, their development is limited by the fact that the book is a material object. Digital technology provides us with new ways of accessing the information contained in books (or any text for that matter) which could help both scholars and laypeople to make better use of published information.

Traditions in the publishing world are however remarkably resilient. The term *incunabulum* refers to the customary appearance of books during the first 45 years of printing in Europe, 1455-1500. Despite the transformations in book production that occurred with the introduction of Gutenberg's printing press, notably the use of movable type in place of the human hand, printers strived to make their printed books as similar to the traditional manuscript books as possible.[3] Gothic script and rubrication was after all what their customers were used to and presumably what pleased the eyes of the printers themselves.[4]

We now have a 500 year history of reading printed books and with the advent of e-publication, publishers are tailoring their digital editions to look like print. Applications like Amazon's *Kindle* reproduce printed pages (pages aren't strictly speaking necessary in an immaterial digital edition) and some applications even include a page-flipping simulation,[5]

[1] The right-hand page on the front cover is taken from the web site: Enhanced edition of *The Book-hunter in London*, <http://bookandbyte.org/bookhunter/showDataPerson.php?person=25> (17 July 2016).

[2] M.A. Rouse and R.H. Rouse, *Authentic Witnesses: Approaches to Medieval Texts and Manuscripts* (Notre Dame, Indiana: University of Notre Dame Press, 1991), pp. 191-219.

[3] L. Hellinga, 'The Gutenberg Revolutions', in S. Eliot and J. Rose (eds.), *A Companion to the History of the Book* (USA, UK, Australia: Blackwell Publishing, 2007), pp. 214-215. 'What are Incunabula?', *Incunabula. Dawn of Western Printing* <http://www.ndl.go.jp/incunabula/e/chapter1/index.html> (18 March 2016).

[4] The Gutenberg Bible was printed using gothic type (also known as blackletter). The printers left gaps for titles and initials which were then handwritten in colour by a rubricator. M.H. Black, 'The Printed Bible', in B.M. Metzger and M.D. Coogan (eds.), *The Oxford Companion to the Bible* (New York and Oxford: Oxford University Press, 1993), p. 611.

[5] Apple's iBooks provides this type of simulation. F. Jabr, 'The Reading Brain in the Digital Age', *Scientific American*, 11 April 2013, n.pag. <http://www.scientificamerican.com/article/reading-paper-screens/> (1 April 2016).

leading some to term these e-books *digital incunabula*.[6] The text is processed with a new technology but instead of taking the greatest possible advantage of it, the publishers imitate the appearance and applications of print.

That's not to say that this is entirely a bad thing. Page numbers are for example a useful way to locate and refer to particular sections of text. It would not be desirable if digital editions replaced printed books, since each have their own strengths and weaknesses. The thesis explores how semantic publications, those that exploit the new digital technologies, can be an *addition* to the printed book and how the combined features of these two methods of publication can in some cases bring out the best result for the user.

When digital editions go beyond the *digital incunabula* they are usually referred to as 'enhanced publications'. The word enhancement can be taken to mean the inclusion of anything other than plain text in a publication: everything from the illumination performed on manuscripts by monks in the Middle-Ages, to images, apps and videos included as 'supplementary material' in modern digital editions.[7] The term 'semantic publication' is used here because the thesis is primarily concerned with *semantic* digital enhancements: enhancements which make the meaning of texts machine-readable and the creation of networks of information which are semantically linked.

Semantic enhancements have already gained a considerable following in the sciences. Leading scientific publishers such as *Springer* and *Elsevier* have submitted formal guidelines for the addition of supplementary data and extensive metadata into their publications,[8] and scientific articles have been the primary subject of most of the writing on digital enhancements.[9] Enhanced publications to a large extent provide the solution to the scientists' problem of data overflow. There are great potential gains from representing articles not merely as electronic PDFs, but making full use of the possibilities of the Semantic Web.[10]

[6] See for example: G. Crane et al., 'Beyond the Digital Incunabula: Modeling the Next Generation of Digital Libraries', in J. Gonzalo et al. (eds.), *Research and Advanced Technology for Digital Libraries*, vol. 4172 (Berlin and Heidelberg: Springer, 2006), pp. 353-366. K. Rowe, 'Living with digital incunables, or a "good-enough" Shakespeare text', in C. Carson and P. Kirwan (eds.), *Shakespeare and the Digital World. Redefining Scholarship and Practice* (UK: Cambridge University Press, 2014), pp. 144-159.

[7] N.W. Jankowski et al., 'Enhancing Scholarly Publications: Developing Hybrid Monographs in the Humanities and Social Sciences', n.pag. <http://ssrn.com/abstract=1982380> (28 April 2016).

[8] D. MacMillan, 'Data Sharing and Discovery: What Librarians Need to Know', *The Journal of Academic Librarianship*, 40:5 (2014), pp. 544-545.

[9] See for example: S. Woutersen-Windhouwer et al., *Enhanced Publications. Linking Publications and Research Data in Digital Repositories* (Amsterdam: Amsterdam University Press, 2009). D. Shotton et al., 'Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article', *PLoS Computational Biology*, 5:4 (2009), n.pag. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663789/> (5 July 2016).

[10] T. Groza, *Advances in Semantic Authoring and Publishing* (Amsterdam: AOS Press, 2012), pp. 3-4.

The concept of the Semantic Web refers to the linking of online data, making connections between digital objects that are related to each other. This involves making these connections machine-readable: the coding has to give explicit commands to a search engine algorithm on how to interpret the connections between the digital objects.[11] The objective is to improve the results returned by search engines. As an example: if all the writings of Charles Dickens are linked to a single entity called ‘Charles Dickens’ which is defined as ‘an author’, the algorithm of a search engine will on command retrieve all the document instances where Dickens is the writer. Without being explicitly told that Charles Dickens wrote these texts, the algorithm can merely do a full-text search of a database and retrieve all occurrences of ‘Charles Dickens’, whether or not that set of digits was mentioned in passing or included on the title page.[12]

The exercise of creating semantic enhancements consists of coding text and weaving it into the Semantic Web. A semantic publication contains semantic links both within the publication and into other online domains. Programming languages use code to make information about a given text machine-readable. XML (the eXtensible Mark-up Language), the programming language most widely used in the publishing world, uses various elements in brackets to relate information to the computer. The element <p> marks the beginning of a paragraph, while </p> marks its conclusion. An example of a code is:

```
<p>It was the best of times, it was the worst of times.</p>
```

This code simply tells the computer: *This is a paragraph.* The <p> element is an example of *metadata*: data which is not a part of the text, but which gives information about the text. The following code goes a step further:

```
<p author=”Charles Dickens”> It was the best of times, it was the worst of times.</p>
```

This code tells the computer: *This is a paragraph and Charles Dickens wrote it.* It informs the computer about more than just structural features. It encodes not only form, but also meaning. This type of metadata is the basis for most of the enhancements discussed in the thesis. The quote would now be searchable as a composition of Charles Dickens.[13]

[11] W3C, ‘Semantic Web Activity’, <https://www.w3.org/2001/sw/> (11 March 2016).

[12] Sophisticated algorithms can make up for the limitations of full-text search to a certain extent, but they can only make educated guesses. One example are so-called ‘proximity operators’: if a search query contains two phrases (‘Charles Dickens AND David Copperfield’) the algorithm gives precedence to texts in which both phrases appear on the same page. J.W. East, ‘Subject Retrieval from Full-Text Databases in the Humanities’, *Libraries and the Academy*, 7:2 (2007), p. 231.

[13] The Semantic Web uses different techniques and codes to make meaning machine-readable (see footnotes 98 and 99, p. 25) but this XML code has the same function.

Most online publications include some metadata, such as information on authorship and circumstances of publication.[14] A semantic publication will however additionally include a substantial amount of metadata encoded into the text itself. The following definition determines the scope of the enhancements explored in this thesis:

> [T]he term *semantic publication* ... include[s] anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between articles.[15]

The only nuance to add is that this definition assumes an article, while our main subject is an academic monograph, a type of publication which is uniquely important in the humanities. We will examine which of the semantic enhancements that have been done in the sciences can be *usefully* replicated in the humanities. The aim is to decipher which type of humanities publication is best suited for an extensive semantic publication, how this publication should be coded, and how readers can eventually make use of the enhancements. There are various stakeholders in the publication of a book: the author, the reader, the publisher etc. Our main focus will be on the reader's interest. The emphasis is on maximising the user-friendliness and usefulness of a semantic publication for its end-user.[16]

While the first chapter of the thesis is concerned with the rationale for making semantic editions in the humanities generally, the second and third chapter examine monograph enhancements through a case study of a single book: *The Book-hunter in London* (1895).[17] The chapters on the design of a semantically enhanced version of this book and on the academic value of these enhancements are intended to demonstrate how a semantic edition of the right type of monograph can bring out the best in digital publishing in the humanities. A semantic edition of the book has been made for the purposes of this project and published on the web.[18]

Finances are always a central consideration when it comes to book publishing and an endeavour like a semantic publication does at some point have to be evaluated in terms of its

[14] For a list of the standard entities of book metadata, see: R. Register, *The Essential Guide to Metadata for Books* (New York: F+W Media, 2013), p. 9.

[15] Shotton et al., 'Adventures in Semantic Publishing', n.pag.

[16] This emphasis is the main reason for omitting the subject of Open Access in the thesis. As pointed out by Agata Mrva-Montoya: '... the open access publishing model ... is driven by experimenting with the new business, distribution and permission models rather than with a new format of scholarly communication practice.' A. Mrva-Montoya, 'Beyond the monograph: Publishing Research for Multimedia and Multiplatform Delivery', *Journal of Scholarly Publishing*, 46:4 (2015), p. 322.

[17] W. Roberts, *The Book-hunter in London. Historical and other Studies of Collectors and Collecting* (Elliot Stock: London, 1895) <http://www.gutenberg.org/ebooks/22607> (12 April 2016) [under 'Case Study' in the bibliography].

[18] Enhanced edition of "The Book-hunter in London", <http://bookandbyte.org/bookhunter/> (31 May 2016).

economic viability. However, before embarking on such an edition, it is ideal to know which of the *potential* enhanced features are actually suited to its end-users. That is the object of this thesis.

The focus throughout is on the *design* and the *use* of the semantic edition rather than the technical implementation of its enhancements. Therefore, no technological knowledge is required of the reader. The thesis is directed towards individuals who have some stake in the publishing of academic texts in the humanities, whether it is as writers, publishers, or readers, and who seek to know more about the possibilities digital technology provides for the consumers of academic books.

Chapter 1: Rationale

> 'And so on, down through the successive decades and generations of the past four centuries, the decline—but not the death, for such a term cannot be applied to any phase of book-collecting—of one particular aspect of the hobby has synchronized with the birth of several others, sometimes more worthy, and at others less.'
>
> W. Roberts, *The Book-hunter in London*, p. 59.

Why should it be worthwhile to do semantic publishing in the humanities in the first place? This chapter will explore the rationale for making semantic enhancements to an academic monograph and identify which types of content would benefit most from a semantic edition.

1. Print and digital

The monograph is commonly defined as 'a printed specialist book-length study of a research based topic', typically based on the research of a single academic. Monographs enjoy a uniquely privileged position as a mode of publication in the humanities, where they are generally viewed as more important than journal articles and are in many cases essential for career advancement.[19] This to some degree goes for the social sciences as well,[20] which is why these two disciplines are often lumped together (short spelling HSS) in discussions of the monograph.[21]

The other dominant form of academic publishing is the journal article. A study comparing citations in the years 1981-2000 in the natural sciences and engineering on one hand, and the social sciences and humanities on the other, found that in the former fields, between 80 and 90% of citations referred to journal articles, while the percentage was between 40 and 50% in HSS.[22] Scholarly output also supports the case for the importance of the monograph. Journal articles represent close to 100% of the scholarly output of the sciences, but substantially less in the humanities. Philosophy comes closest to the sciences, with 60% of its output as journal articles.[23] Another study found that among academics, humanists are by far the most avid readers of books, with the social sciences coming in second and engineering at a distant third.[24] The prestige of the monograph is not only symbolic, but also reflected in usage.

[19] P. Williams et al., 'The role and future of the monograph in arts and humanities research', *Aslib Proceedings*, 61:1 (2009), p. 67.

[20] G. Crossick, 'Why Monographs Matter', n.pag. [under 'Unpublished secondary sources' in bibliography].

[21] For example: V. Larivière et al., 'The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities', *Journal of the American Society for Information Science and Technology*, 57:8 (2006), pp. 997-1004.

[22] Ibid., pp. 1000-1003.

[23] Crossick, 'Why Monographs Matter', n.pag.

[24] C. Tenopir, R. Volentine and D.W. King, 'Article and book reading patterns of scholars', *Learned Publishing*, 24:4 (2012), p. 287.

Printed books and e-books each have unique qualities. This chapter will argue that in the case of the academic monograph, some publications could benefit from a careful combination of print and digital publication. Rather than focusing on what we might call emotional preferences, such as the physical size or the touch of a book,[25] the chapter will focus exclusively on practical aspects of the academic reading experience. Which features should be enhanced to maximize the user-friendliness of an academic monograph and which features of print and digital are most relevant to this goal?

Keeping the end-user in mind, we must first consider how the average consumer of the academic monograph wants to read. One of its primary target groups, college students, is heavily biased towards print. A 2010 study found that aside from the students preferring printed books to e-books, 'previous experience with e-books [did] not increase preference for e-books', and this despite the students' frequent computer use.[26] According to studies by the American linguist Naomi S. Baron, today's American students prefer printed books in all categories of publication (aside from academic journal articles, which are often only accessible on the web). Print is considerably more popular than digital both in the students' reading for school and for pleasure.[27] There are also indications that the growth of the eBook's market share in publishing is slowing down.[28] According to a 2011 report by the UK's *Research Information Network,* humanities scholars still favour libraries over web based products.[29] At the very least, printed books are not on the way out anytime soon.

And when it comes to the academic monograph, not all the advantages belong to the digital book. Printed books have several qualities which are essential to the academic monograph. Unlike an online publication, a printed book is permanent and unchangeable. Once it has been published, the text is fixed and cannot be altered. This is commonly referred to as the *fixity* of print. An online publication can however easily be tampered with after

[25] T. Blanke et al., 'Digital Publishing Seen from the Digital Humanities', *Logos,* 25:2 (2014), p. 18. These features are a topic in and of themselves and may partly explain the general preference for printed books.

[26] W.D. Woody, D.B. Daniel and C.A. Baker, 'E-books or textbooks: Students prefer textbooks', *Computers & Education*, 55:3 (2010), p. 947.

[27] N.S. Baron, *Words Onscreen. The Fate of Reading in a Digital World* (Oxford and elsewhere: Oxford University Press, 2015), pp. 83-84.

[28] T. Tivnan, 'E-book sales abate for Big Five', 29 January 2016, n.pag. <http://www.thebookseller.com/blogs/e-book-sales-abate-big-five-321245> (19 March 2016). M. Bluestone, 'AAP StatShot: Publisher Net Revenue from Book Sales Declines 4.1% in First Half of 2015', 8 October 2015, n.pag. <http://publishers.org/news/aap-statshot-publisher-net-revenue-book-sales-declines-41-first-half-2015> (19 March 2016).

[29] *Reinventing research? Information practices in the humanities* (UK: The Research Information Network, 2011), p. 6 <http://www.rin.ac.uk/system/files/attachments/Humanities_Case_Studies_for_screen_2_0.pdf> (5 March 2016).

publication, which makes citation more problematic.[30] A printed book will also not pop out of existence suddenly. This may seem unremarkable, but e-books are not permanent in this sense. Their accessibility depends on someone paying for their presence on an online server. If the e-book is no longer hosted on the server, the access is gone.[31]

The fixity of the printed version is in some ways more vital to the humanities than the sciences, given that research in subjects like history tends to stay relevant for longer than most scientific research. In the light of the unstable nature of online publications, a humanities monograph needs to be rooted in the permanence and fixity of the printed book.[32] Rather than being seen as an exclusive publication, a semantic edition of a humanities monograph should therefore be seen as an extension of the printed book. The printed book should be fully usable independently of its semantic counterpart version, which is subject to change and could disappear altogether.[33]

Research on students' use of textbooks and e-books suggests that students make less use of special features, such as charts, in digital editions than in print.[34] While one should be wary of drawing too extensive conclusions from this, the results nevertheless indicate that a semantic edition should be focused on elements which cannot be reproduced in a printed book, rather than replicating features which work superior in print. Printed books are for instance user-friendlier in terms of annotation. The *Research Information Network*'s report did cite 'inadequate annotation tools' as a barrier to the use of online resources by humanities scholars.[35] However, humans are not as nimble with a computer mouse as with their fingers, so online annotation will probably never match the spontaneity and intricate mind-mapping allowed by a pencil. Since the primary users of humanities *monographs* (as opposed to other online resources) seem more interested in consulting printed copies, it is safe to assume that they will want their annotations there on the page as well.

Browsing a book through page-flipping is another feature which works better in a printed book. In the book-length argument typical of humanities monographs, the reader will often want to refer to earlier pages, flip back and forth, and this is far faster and more efficient

[30] A. Van der Weel, *Changing our textual minds. Towards a digital order of knowledge* (Manchester and New York: Manchester University Press, 2011), p. 149-150.
[31] Ibid., p. 145.
[32] Williams et al., 'The role and future of the monograph', p. 78.
[33] To be precise, the printed book's importance in the sense discussed here does not fundamentally rely on it being a printed object, but rather on the fact that it presents a stable and definitive version of a text in a way which an online publication cannot. The printed book is stable and fixed because it is an *analogue* object. An analogue presentation of text in any format would fulfil the same function. The printed book simply happens to be a user-friendly and culturally recognized object for carrying out this task.
[34] Woody, Daniel and Baker, 'E-books or textbooks', p. 947.
[35] *Reinventing Research?*, p. 73.

using a printed book. While page-flipping might not be a strong point of digital editions, a more direct search of a book's content with the help of a search engine is a feature of the Web which the printed version cannot replicate. The search capacity available in a digital corpora of texts is both quicker and more efficient,[36] and has for instance resulted in the widespread digitisation of dictionaries, which are exclusively used for precise topic queries. This discoverability of topics is one of the major advantages of a semantic digital edition and will be explored further in the second chapter.

Another way to view this browsing advantage of the digital edition is to say that it is more favourable to non-linear reading than a printed edition. Rather than reading a book from cover to cover, a user can search its content based on keywords, and is therefore more likely to look only at the pages which contain material directly relevant to his or her subject of interest.[37] Prime candidates for extensive digital editions are therefore books which are highly likely to be used in this way: ones that include a lot of descriptive content, cover a vast amount of different topics (within an overarching theme) and where each chapter is to a great degree self-contained.

In this light, books that aim to give an overview of a subject, such as educational textbooks or a collection of chapters on a single theme, seem the most obvious choice for a semantic edition. Meanwhile, a book which is structured as a linear narrative, such as an autobiography, would in these terms benefit less from a semantic counterpart. An autobiography's topic is restricted to a single person and biographies tend to explain a person by taking their entire personal history into account: fully understanding one chapter depends on having read the previous ones.

A factor that is unique to digital editions is the possibility for multimedia applications. In addition to text, audio, video, games and links are available in the digital sphere.[38] While these provide many opportunities, multimedia can also be seen as a deficiency of the digital world. As pointed out by the publisher and writer Michael Bhaskar: 'Units of attention represented by the book remain consistent. Infinite content and hyperlinking there may be, infinite attention there is not.'[39] Academic monographs in the humanities are often essentially book-length arguments, so the distractions of online multimedia can be damaging to a sustained attention to the

[36] J.B. Thompson, *Books in the Digital Age. The Transformation of Academic and Higher Education Publishing in Britain and the United States* (UK and USA: Polity Press, 2005), p. 319-320.

[37] The scholar Terje Hillesund has called this reading behaviour 'fragmented reading'. His research showed that academics prefer to do 'concentrated reading' on paper, but tend to skim web pages. T. Hillesund, 'Digital reading spaces: How expert readers handle books, the Web and electronic text', *First Monday*, 15:4 (2010), n.pag. <http://firstmonday.org/article/view/2762/2504> (23 May 2016).

[38] Van der Weel, *Changing our textual minds*, pp. 153-154.

[39] M. Bhaskar, *The Content Machine. Towards a Theory of Publishing from the Printing Press to the Digital Network* (UK and USA: Anthem Press, 2013), p. 50.

development of that argument. This distractive nature of multimedia, combined with the academic readers' preference for print, suggests that multimedia options should be available to readers *when they consult them specifically*. The reader can break away from the printed edition to use the semantic edition.

Of course, varying levels of digital enhancements will be appropriate for different titles. While we are primarily concerned with providing the most useful digital features for the right type of humanities publication, some thought has to be given to financial matters. Aside from a monograph having the characteristics previously mentioned, an extensive semantic edition will not be made unless there is a relatively large audience for the book, and that the book is likely to be in use for some time in the future. But even though a highly specialized monograph with a print run of 200-300 copies does not call for an elaborate digital edition, the more basic levels of semantic enhancement are equally useful for these books. They receive the same benefit from the fixity of print and online discoverability. In a world of digitised scholarship, all printed monographs are in need of some level of digitisation. These digital enhancements should be seen as an extension of a stable and permanent text, which should be fully independent of its digital counterpart.

2. Fundamental differences in research practices

The intended audience of a publication determines which features it should have. To decide which features of the semantic publications that have been made in the sciences are useful in the humanities, it is instructive to look at some fundamental differences in these two broad areas of research.

The first thing to note is that the division into research areas is to an extent semantic and varies between languages. The linguist Anna Wierzbicka contrasts the English definition of the word 'science', which is strictly separated from the humanities, math and logic, with the German term 'Wissenschaft', which is an umbrella term for all knowledge accumulated in a systematic way.[40] She traces the roots of the English distinction between sciences, social sciences, and the humanities to the Italian eighteenth century philosopher Giambatista Vico. The base of the distinction is that, as is inherent in the word, the humanities study people, not things. Things are the object of science. The social sciences apply an empirical scientific investigation to groups of people, studying them in the same manner the sciences would study things. This leaves the non-empirically driven research into peoples' existence to the humanities.[41]

[40] A. Wierzbicka, 'Defining "the humanities"', *Culture & Psychology* 17:1 (2011), p. 33.

[41] Ibid., pp. 34-35.

While the arguments of humanities scholars are generally not centred on empirically testable propositions, there is nevertheless often a multitude of elements within a humanistic study which *can* be studied empirically. Let's take for example the question of why Vincent van Gogh cut off his ear. There are qualitative and quantitative ways of approaching this question. Qualitatively, we can look at studies of Van Gogh's life and try to trace his mental development from the close reading[42] of existent primary and secondary sources. Quantitatively, we could text mine[43] the correspondence between Vincent and his brother Theo. We could for instance investigate which words appear most often, and see what the results reveal about Van Gogh's mental state.

However, neither of these measures will ever give us a precise or particularly secure understanding of what was happening in Van Gogh's mind in the time before he cut off his ear. This evidence is very indirect, at least in the scientific sense. If we had direct access to Vincent, we could give him a sociological questionnaire. We could scan his brain and analyse it according to our current understanding of that organism. Using the analogy of Wierzbicka, we could study him as a 'thing'. In the absence of Vincent himself, our subject, we can't do any direct testing. We're bound to study him as a 'person'.

While these limitations don't apply to all studies in the humanities, I believe they provide clues to why the monograph has such a central role within the humanities. As noted by Geoffrey Crossick, the argumentation presented in an academic monograph is 'of a specific character that ... cannot be replicated or modelled, [which] means that there is a need to present thick description and more direct evidence'.[44] The phrase 'more direct evidence' can be taken to mean *less abstract*. Studies in the humanities rely more heavily on qualitative data than studies in the sciences: humanists tend to look directly at their sources, rather than taking a step back and examining them statistically. So studies in the humanities are indirect in the sense of not engaging directly with their physical objects of study, people; but direct in the sense of examining the evidence closely, rather than observing it through the lens of statistics.

This observation sits well with Wierzbicka's definition of the humanities, that they are not centred on empirical investigation. The humanities require 'thick description', detailed arguments about various perspectives of the subject, because the subject is often not tangible,

[42] As opposed to *distant reading*: the study of the formal aspects of a large corpora of texts with the aid of an algorithm. Close reading simply refers to the traditional practice of individual reading. See F. Moretti, *Distant reading* (London and New York: Verso, 2013).
[43] Text mining is the practice of applying algorithms to a collection of digitised texts in order to retrieve information on some abstract feature of these texts.
[44] G. Crossick, *Monographs and open access. A report to HEFCE* (London: HEFCE, 2015), p. 13-14 <http://www.hefce.ac.uk/media/hefce/content/pubs/indirreports/2015/Monographs,and,open,access/2014_monographs.pdf> (11 March 2016).

such as in philosophy, or available for direct testing, such as when dealing with historical figures or analysing the deceased authors of literature.

Therefore, many different aspects of a single subject come together in a humanities monograph. The larger context becomes disproportionately important in comparison with the sciences, and this has consequences for a semantic publication. In a way, the jungle of information which the humanities scholar has to wade through is much larger than that of the sciences. Not necessarily in terms of bibliographic material, but simply the sheer breadth of their object of study: human beings. The humanities are concerned with the entire history of humanity, its art and rituals, its written record and history, its poetry and literature. To find their way through this jungle the humanities scholar needs to understand the large context and a semantic edition can help him to do this.

The monograph provides the opportunity for humanities scholars to '[embed] their research in a larger scholarly, temporal and spatial network'.[45] The semantic publication could be seen as a digital representation of this network. Mapping out the context in which the main subject is depicted is the central issue in a semantic edition of a humanities monograph. The enhancements should help to clarify connections within a given book and between different publications.

Another feature which distinguishes the humanities from the sciences is that there is generally less large-scale collaboration in HSS studies. This may be due to inherent differences in the type of knowledge the sciences and the humanities respectively seek. The biologist Klaus Jaffe has applied methods previously used in research on ant behaviour, to research strategies in the academic world. The ant research used computer simulation to study how the foraging strategies among the ants differed on the basis of their 'resource landscape', whether their resources were dispersed widely or concentrated in few places. Jaffe employed this same simulation to study different academic fields depending on their 'knowledge landscapes', whether research in these fields was concentrated or dispersed.[46]

The simulation tested two types of strategies which turned out to distinguish a clear difference between the natural sciences on the one hand, and the social sciences and humanities on the other. The first strategy was the 'Democratic system', 'w[h]ere [sic] all workers eventually perform all tasks' and where 'the first discovery will draw the most recrutees

[45] Ibid., p. 14.

[46] K. Jaffe, 'Social and Natural Sciences Differ in Their Research Strategies, Adapted to Work for Different Knowledge Landscapes', *PloS one*, 9:11 (2014), n.pag. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113901> (13 June 2016).

[sic]'.[47] This system turned out to be ideal for a knowledge landscape which consisted of a 'few large knowledge clusters'.[48] The natural sciences, where the emphasis is on 'a few general basic problems that are the same everywhere',[49] largely conform to this system. Originality is less important than accumulative research, following-up on what other investigators are doing is essential.[50]

The second strategy is the 'Technocratic system', 'where workers specialize either in scouting or in retrieval and w[h]ere [sic] the society collects several smaller resources simultaneously'.[51] This system is ideal in the social sciences and humanities, where the clusters of knowledge are many, small and researchers in different sub-areas are largely working in isolation of each other.[52] This is also reflected in publication. On average, the natural sciences publish in few journals with high citation rates, while the social sciences and humanities publish in many journals, with fewer articles and fewer citations.[53]

This focus on originality in the humanities could be reflected in the choice of which texts are to be semantically enhanced. While follow-up research is vital to the sciences, an academic in the humanities is less likely to want to follow-up on a topic which has already been examined in a monograph, particularly since many monographs are quite highly specialized. For example, the writer of *Shakespeare and the Renaissance Concept of Honor*[54] has probably exhausted the research interest in that particular topic. The ideal semantic publication should have a more general theme and act like a crossroad between various different perspectives on that theme. It should spread its digital tentacles as widely as possible, reaching into a variety of different knowledge relevant to the monograph's subject, hopefully stimulating the reader to discover new terrains to explore.

3. Bibliographic Metadata

Science is to a large extent a data-driven enterprise and scientific papers are consequently a prime candidate for enhanced publishing. It is therefore hardly surprising that when enhanced

[47] Ibid., n.pag [under the heading 'The Model'].
[48] Ibid., n.pag [under the heading 'Simulations'].
[49] Ibid., n.pag [under the heading 'Empirical bibliographic evidence'].
[50] Ibid., n.pag [under the heading 'Discussion'].
[51] Ibid., n.pag [under the heading 'The Model'].
[52] Ibid., n.pag [under the heading 'Discussion'].
[53] Ibid., n.pag [under the heading 'Empirical bibliographic evidence'].
[54] C.B. Watson, *Shakespeare and the Renaissance Concept of Honor* (United States: Princeton University Press, 2015).

articles connected to the web were first being discussed in 2001, they were primarily presumed to benefit the sciences.[55]

In 2007, Michael Seringhaus and Mark Gerstein pointed out in an article on molecular biology that this data-driven discipline could benefit greatly from direct access to relevant data through its journal articles. They suggested that supplementary data should be handed in along with the text of an academic publication, that articles should be 'fully computer-readable with intelligent markup', and that all relevant external data, such as 'textbooks, laboratory Web sites and high-level commentary' should be adequately linked to the articles. Furthermore, the enhanced publications should have functions for peer-review and other commentary, and all the different articles on biology should be searchable through a single portal.[56] The sciences have already to a great extent fulfilled these promises, with many scientific publishers demanding data to be handed in along with journal articles,[57] and the existence of databases such as *PubMed Central*, 'a free full-text archive of biomedical and life sciences journal literature' with 3.8 million articles available through a single portal.[58]

Meanwhile, a number of studies indicate that academics in the social sciences and humanities (HSS) are a lot less enthusiastic in their use of digital tools.[59] While the sciences have been at the forefront of all levels of semantic publishing, the lag of the HSS fields is particularly noticeable with regard to the most basic function of semantic publications, that is: to '[facilitate] … automated discovery'.[60]

Making published texts visible on the Internet is important because the academic world has moved online. The Internet is one of the most common means of finding information in both the physical and life sciences.[61] And even though HSS scholars have a greater fondness for

[55] T. Berners-Lee and J. Handler, 'Publishing on the semantic web', *Nature,* 410:26 (2001), pp. 1023-1024.

[56] M.R. Seringhaus and M.B. Gerstein, 'Publishing perishing? Towards tomorrow's information architecture', *BMC Bioinformatics*, 8:17 (2007), n.pag. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-17> (12 March 2016).

[57] MacMillan, 'Data Sharing and Discovery', pp. 544-545.

[58] PubMed Central, <http://www.ncbi.nlm.nih.gov/pmc/> (13 March 2016).

[59] *Reinventing research?*, p. 70.

[60] Shotton et al., 'Adventures in Semantic Publishing', n.pag. In fact, this appears not to be a recent development. Back in 1983, Stephen Wiberley complains that 'machine-readable systems for information retrieval in the humanities … have lagged behind the systems available for the sciences and social sciences.' S.E. Wiberley, Jr., 'Subject Access in the Humanities and the Precision of the Humanist's Vocabulary', *The Library Quarterly,* 53:4 (1983), p. 432.

[61] *Collaborative yet independent: Information practices in the physical sciences* (UK: The Research Information, 2011), p. 69 <http://www.rin.ac.uk/system/files/attachments/Phys_Sci_case_study_full_report.pdf> (8 June 2016). *Patterns of information use and exchange: case studies of researchers in the life sciences* (UK: The Research Information Network, 2009), p. 36 <http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf> (8 June 2016).

physical libraries,[62] they browse the Internet rather than the library bookshelves to locate information.[63]

It is already a standard to make humanities texts machine-readable. Each letter is represented in the text through a binary code,[64] as opposed to a non-searchable text on an image, making an online full-text search possible.[65] Even this move from the analogue book to the *digital incunabula* does not permit the computer insight into the *meaning* of words in a text, which often depends on their context. An algorithm doing a full-text search could for instance not distinguish between 'When the bird *leaves* its nest' and 'The autumn *leaves* of red and gold'. Without some form of metadata for guidance, a search engine is much less efficient at retrieving the most relevant results to a query. The addition of metadata is therefore an important factor in making sure a text reaches its audience.

Two types of metadata are most relevant to the discoverability of an online publication. These are distinguished by the level at which they are encoded. The type of metadata which is intended to identify the book and its content is called *bibliographic* metadata.[66] This is the metadata which search engines make use of to find digital objects (such as books and articles) and includes information about circumstances of publication, and keywords for topics. Bibliographic metadata is encoded at the level of the individual object. The other type of metadata is encoded directly into the text of a publication and is intended to help the reader find information *within* the text. This will be discussed in the second chapter.

The emphasis on discoverability in the sciences is exemplified by an article published in 2015 in the *Journal of the Medical Library Association,* which made obligatory the admission of a specific set of keywords along with any published article. The article stated that the additional visibility 'should improve journal visibility, subsequent citation counts, and its impact'. The title of the article, the abstract and the keywords, together should make up a 'miniaturized version of [the] paper'.[67]

On the whole, bibliographic metadata does seem to be making major strides in all areas of academic publishing, including the humanities. In 2015, the *Online Computer Library Center,* an international library association which curates the well-known *WorldCat* database,

[62] *Reinventing research?,* p. 6.
[63] Ibid., pp. 24-25.
[64] At the most basic level of computation, computers only register a series of 0 and 1, a binary calculation.
[65] Van der Weel, *Changing our textual minds,* p. 146.
[66] See Hathi Trust Digital Library, 'Bibliographic Metadata Specifications', <https://www.hathitrust.org/bib_specifications> (28 April 2016).
[67] T. Bekhuis, 'Keywords, discoverability, and impact', *Journal of the Medical Library Association,* 103:2 (2015), p. 119.

made deals with major publishers in business, social sciences and the humanities to make their publications discoverable online. This metadata extended to 'books, e-books, journals, audio-visual materials and databases'.[68]

Despite the importance of a book's online presence, the addition of bibliographic metadata to publications in the humanities is nevertheless in some ways being neglected. In most databases specialized in the humanities, a multi-authored book will have specific metadata for all of its chapters. However, a single-author monograph is frequently only given the same amount of metadata as a single article, despite containing much more data.[69]

As a consequence of the lack of online visibility, a lot of useful material may never find its audience. While the generally short papers published in the sciences tend to have more than ten keywords assigned to them, a study on monographs in the area of philosophy found that each had only 5.6 subject terms (a type of bibliographic metadata) on average assigned to it. This turned out to be a subject term every 48 pages on average. In the OPAC catalogue (Open Public Access Catalogue), the number of subject terms associated with a monograph turned out to be 3.1 on average, or one for every 88 pages.[70]

And while the aforementioned study is ten years old, a look at the *WorldCat* catalogue and the *Leiden University Library Catalogue* indicates that little has changed. In the *WorldCat* database, if one requests all English non-fiction books with the subject 'houdini', published between 1960 and 1975, the average number of subjects per book is 5,6. If one book with an exceptionally high number of subject terms (31) is excluded, the average falls to 4,3.[71] A great number of the titles only list the name of the magician in various versions. The same query for the years 2010-2015, produces an average of 6,7 subjects per book,[72] not a very substantial increase. In the latter period there is however generally a lot more information in the way of summaries, abstracts and chapter titles.

Bibliographic metadata can be implemented with relatively little effort but the rewards are very significant. To name just one example, the book *Nature and Love in the Late Middle Ages* (1963) has three subject terms assigned to it in the Leiden University library catalogue, and 5 in the *WorldCat* catalogue.[73] The period defined in the title makes it unlikely that a student of

[68] 'OCLC signs agreements with publishers in the Humanities, Social Sciences and Business', OCLC, 14 April 2015, n.pag. <https://www.oclc.org/en-CA/news/releases/2015/201512dublin.html> (28 February 2016).

[69] J.W. East, 'Subject retrieval of scholarly monographs via electronic databases', *Journal of Documentation*, 62:5 (2006), p. 599.

[70] Ibid., pp. 599-600.

[71] WorldCat, 'Advanced search' <https://www.worldcat.org/advancedsearch> (16 March 2016). Subject: 'houdini'. Year: 1960-1975. Audience: Non-juvenile. Content: Non-fiction. Format: Book. Language: English.

[72] Of the 28 results, two were excluded because they dealt with the *Houdini* software, rather than the magician.

[73] The web sites are <http://catalogue.leidenuniv.nl/> and <https://www.worldcat.org/>.

the Enlightenment thinker Jean-Jacques Rousseau will come across a substantial chapter comparing the naturalism of the late-medieval period to Rousseau's naturalism.[74] If every chapter in the book were lavished with the same amount of metadata as the average scientific journal article, the search engine would not miss out on this discussion.

Making humanities monographs discoverable is all the more important in light of their declining sales. Monographs by now often have a print run of as few as 200-300 copies,[75] and are therefore unlikely to be available in print to those who could benefit from them at their local library. Their users will probably not encounter them strolling between the library bookshelves or even in an academic bookstore, few of which remain. Research has shown that scholars in the humanities tend to shy away from bibliographic databases, using services such as the Amazon recommendations and Google Books to find their sources. One of the hindrances to the scholars' use of these databases turned out to be the focus on journal articles. Bibliographic information on monographs tended to be left out of the records.[76]

Bibliographic metadata is a basic enhancement relevant to every monograph published, no matter how small the audience, or perhaps even more so when the audience is very small. Publications have to be made visible to their potential users, who mostly do their book-hunting online.

4. Data-intensity and the digital humanities

Enthusiasts in the area of semantic publishing have stated that '[s]cientific innovation depends on finding, integrating, and re-using the products of previous research'.[77] According to studies done by the *Research Information Network* in the UK in 2009 and 2011, researchers in both the life sciences and the physical sciences are making substantial use of digital technology to meet these ends. Collections of data in online repositories are considered a 'new paradigm in the life sciences'[78] and physical libraries are on the way out in this field.[79] The report on the physical sciences concludes that they are '[in] many ways ... at the forefront of using digital tools and methods to work with information and data'.[80] When users have 'access to data within [an] article in actionable form' they can verify themselves whether they think the data is valid and

[74] A.D. Scaglione, *Nature and Love in the Late Middle Ages* (Berkeley and Los Angeles: University of California Press, 1963), pp. 136-144.

[75] Williams et al., 'The role and future of the monograph', p. 69.

[76] Ibid., p. 76.

[77] Shotton et al., 'Adventures in Semantic Publishing', n.pag.

[78] *Patterns of information use and exchange*, pp. 8-9.

[79] Ibid., p. 36.

[80] *Collaborative yet independent*, p. 4.

make their own observations about it. Data-sharing has been one of the main preoccupations of scientific publishing in later years.[81]

The social sciences and humanities have not followed the sciences in their emphasis on sharing and re-using data.[82] It has been suggested that the data-intensity of the scientific disciplines is what has led them to expand much faster into online data-sharing than the humanities disciplines.[83] Research in the humanities, as it is done today, may in fact be less data-intensive, but should it be so? Proponents of the digital humanities favour a greater emphasis on data in the humanities. The academics involved have hotly debated the exact definition of the discipline,[84] but an important part of it is the humanities scholars' creation of their own data.

All text is data. The plain text of a monograph can be categorised as *unstructured data*: data 'in which the boundaries of individual items, the relations between items, and the meaning of items, are mostly implicit'. The supplementary data of scientific research and the output of digital humanities projects is however typically categorised as *structured data.* The results are commonly a database 'in which all key/value pairs have identifiers and clear relations and which follow an explicit data model'.[85]

This is a novel form of output for the humanities. As noted by Michael Bhaskar:

> The growth of new disciplines like the digital humanities, whose outputs are data sets, websites or software, challenges the monograph and by extension the edifice of scholarly publishing ... Suddenly the fusty academic press has no choice but to introduce products utterly alien to the old enterprise.[86]

Another way of looking at the issue is to say that the digital humanities call for enhanced editions. Rather than viewing print and digital as enemies, where one format is challenging the other and where one's gain is the other's loss, a combined edition in print and digital can be seen as appropriate for some publications. As we've seen, today's students still like to read books in print, but they also belong to a tech-savvy generation.

The digital humanities have been gathering pace in the last few years,[87] and if future research in the humanities will become to a larger extent data-driven, semantic editions can be

[81] See for example: A. de Waard, 'The future of the journal? Integrating research data with scientific discourse', *Logos*, 21:1 (2010), pp. 7-11. One example of a scientific data sharing project is the *GenBank*, 'an annotated collection of all publicly available DNA sequences'. NCBI GenBank, 'GenBank Overview', <http://www.ncbi.nlm.nih.gov/genbank/> (7 May 2016).

[82] *Reinventing research*, p. 74.

[83] MacMillan, 'Data Sharing and Discovery', p. 542.

[84] See M. Kirschenbaum, 'What is Digital Humanities and What's It doing in English Departments?', in M.K. Gold (ed.), *Debates in the Digital Humanities* (Minneapolis and London: University of Minnesota Press, 2012), pp. 3-11.

[85] C. Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities* 2:3 (2013), n.pag, <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> (2 April 2016).

[86] Bhaskar, *The Content Machine,* pp. 52-53.

of value for publishing the results. Data charts can of course be printed into analogue books, but interactive or actionable data are one of the multimedia options which are unique to digital editions. In a humanities monograph whose argument to some extent centres on structured data, this material will be equally useful as in a scientific journal article, among other things for the reader to verify the assumptions behind the creation of the data.

As previously mentioned, the maintenance of the enhanced version of a book online is dependent on someone hosting it on a server. That combined with the ever evolving online applications, which make the digital editions vulnerable,[88] means that the definitive printed version of the book should be fully understandable independently of the interactive data.

Conclusion

Semantic enhancements provide several advantages for humanities monographs which printed versions alone cannot accomplish. Although printed books are better for tasks such as annotation and page-flipping, they don't allow for the quick and efficient browsing which the semantic enhancements provide.

The semantic version provides valuable features beyond improved subject access. Context is all important in many fields of the humanities, and a semantic version of a printed monograph clarifies the context of the book's content. Since the users of search engines are looking for specific material within a book, rather than reading it linearly, publications which favour non-linear reading benefit most from this contextualisation of data, books in which each chapter is to a degree self-contained.

There is more of an emphasis on original research topics in the humanities than in the sciences, and the digital network provided by the semantic edition can help researchers discover new areas of interest using the semantic links. Publications in the growing discipline of the digital humanities can also benefit from the multimedia possibilities of an enhanced publication.

Different monographs benefit from different degrees of enhancements. Since people browse for books online, bibliographic metadata is relevant to all publications. The more elaborate enhancements mentioned above need more intricate coding to be implemented. That is the subject of the second chapter of the thesis.

[87] Kirschenbaum, 'What is Digital Humanities', p. 9. R. Grusin, 'The Dark Side of the Digital Humanities: Dispatches from Two Recent MLA Conventions', *Journal of Feminist Cultural Studies* 25:1 (2014), p. 82.

[88] D.V. Pitti, 'Designing Sustainable Projects and Publications', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, UK and Australia: Blackwell Publishing, 2004), pp. 472-473.

Chapter 2: Coding & Weaving

'It is infinitely easier to name those who have collected books in this vast and unwieldy London of ours, than it is to classify them. To adopt botanical phraseology, the *genus* is defined in a word or two, but the species, the varieties, the hybrids, and the seedlings, how varied and impossible their classification!'

W. Roberts, *The Book-hunter in London*, p. xvi.

When a decision has been made to add extensive semantic encoding to a monograph in the area of the humanities, how should it be coded and what are the problems and advantages the humanities have in this regard in comparison with the sciences? How could the publication be woven into the Semantic Web?

1. The Case study: *The Book-hunter in London*

The case study that has been selected for semantic enrichment is *The Book-hunter in London*, published by the publishing company *Elliot Stock* in 1895.[89] The writer was William Roberts (1862-1940), an expert on British art who worked for *The Times* as an art critic and art sales correspondent. Aside from writing on British art, Roberts also authored books on the history of bookmaking and book-collecting. Roberts was an ambitious cataloguer of sales records[90] and this passion for gathering information on the minute details of the book trade shines through in *The Book-hunter in London.* There is an overflow of information on prices of books and ownership of collections. The title could have been autobiographical.

The book is a meticulously researched historical work and provides an entertaining, though occasionally overly precise, account of books on the London market and the peculiarities of their collectors over the course of history. The author claims to have taken '[t]he greatest possible care ... to prevent inaccuracy of any kind'[91] and the book includes a lot of precise data to support this claim. The subject matter has a clear relevance to the study of book history.

[89] W. Roberts, *The Book-hunter in London. Historical and other Studies of Collectors and Collecting* (Elliot Stock: London, 1895) <http://www.gutenberg.org/ebooks/22607> (12 April 2016) [under 'Case Study' in the bibliography]. Elliot Stock (1838-1911) had a long career in publishing books and magazines through his own publishing company (active from 1859-1939). Stock initially focused on religious material, but in the 1880's and 1890's started putting out publications primarily of interest to antiquarians and bibliographers. London's most prominent bibliophiles used to meet up in the reading room of the company's office at 62 Paternoster Row, London. *Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland* (Gent and London: Academia Press and the British Library, 2009), p. 198.

[90] Paul Mellon Centre, 'William Roberts', <http://www.paul-mellon-centre.ac.uk/collections/archive-collections/william-roberts> (12 April 2016). B. Allen, 'Paul Mellon and Scholarship in the History of British Art', in *Paul Mellon's Legacy. A Passion for British Art. Masterpieces from the Yale Center for British Art* (New Haven and London: Yale University Press, 2007), p. 45.

[91] Roberts, *The Book-Hunter in London*, p. xiv.

A text in the area of history has been selected for various reasons. To start with, history is a very interdisciplinary field and intersects with many other disciplines of the humanities, including literary studies, sociology, archaeology and philosophy. It is arguably the humanities field which relies most heavily on real-world context,[92]examining a large roster of sources and reflecting on how they can support a general interpretation, which as we've seen is an exercise in which a semantic edition can be of help.

Aside from the fact that the chosen book needs to be a digitised version of a paper-based original and free of any copyright restrictions, a historical primary source such as *The Book-hunter in London* is ideal in other ways. Research already exists on the enhancement of relatively recent university textbooks and anthologies in the humanities and social sciences.[93] This case study provides an opportunity to explore many of the dilemmas of designing semantic enhancements which are generally less prominent in newer books and are particularly relevant to the humanities. In more recent publications, the terminology used is clearer to us and the writer might be present to give an interpretation of his own work. Questions such as how to interpret the author's writing, how precisely to mould the data modelling to each individual book, and how to deal with ambiguous information become more pertinent in an edition of an old source text.

The Book-hunter in London also has many of the ideal features for semantic enhancements discussed in the previous chapter. It has a theme which is relatively general: book-hunting in London throughout history. It provides a historical overview of the subject matter as well as several specific viewpoints on it, such as the chapter 'Women as book-collectors'. The author's interests within this area of study are so wide and varied that they allow each reader to explore a particular niche they are interested in, thus supporting the generation of the individualised original research questions typical of the humanities.

The book also favours non-linear reading, both in the sense that the chapters can be read independently of one another, and with regard to the wealth of descriptive data which it contains. The encoding of this data provides opportunities for retrieving bits of information from the text to support a variety of research with a relation to the general theme of the book.

The text naturally does not support all potential enhancements: The book is rather anecdotal in style and there is no book-length argument. The text is originally printed rather

[92] As observed in a survey on semantic technologies for the study of history: 'Historical data are extremely context dependent, and always open to a variety of possible interpretations.' A. Meroño-Peñuela et al., 'Semantic Technologies for Historical Research: A Survey', *Semantic Web*, 6:6 (2014), n.pag. <http://content.iospress.com/articles/semantic-web/sw158> (28 April 2016).

[93] N.W. Jankowski et al., 'Enhancing Scholarly Publications', n.pag.

than handwritten. When dealing with handwritten texts in multiple editions, semantic issues relevant to the humanities arise, such as how to present varying versions of the same text. *The Book-hunter* does nevertheless have all the most necessary qualifications for a case study on semantic enrichment.

The case study was developed under the pretension that there was a sizeable audience interested in the book. Whether that audience exists in the real world is not the point, the thesis is not concerned with the project's commercial viability. Rather, it is concerned with the *intellectual* viability of enhancements *like* these on a project *like* this and how these enhancements could be accomplished using semantic technologies.

2. Coding: Designing a database

2.1 Databases and ontologies

While digital text is a relatively recent phenomenon, the English professor Martin Mueller points out that text technology goes back a long way in the history of writing. In fact, it could be contended that '[m]edieval monks were the first to turn a text into a database'. Recognizing that human memory was incapable of containing all the different verses of the Bible, the monks divided it into verses and created an alphabetized index. This gave readers the ability to find all the information on a specific subject in one place, and thereby appreciate the harmony of God's word. This system, referred to as the *Biblical concordance*, provided the framework for centuries of analogue database work.[94]

The digital database is engaged in more or less the same task. The creation of a database still relies on carefully crafted indexing systems and taxonomies, a scholarly tradition which has its root in the Middle-Ages. A database is simply 'a system that allows for the efficient storage and retrieval of information'.[95] The *Biblical concordance* defines subjects and then groups all occurrences of it under a single heading in the index. Put more abstractly, it defines entities within the text and the relationship between these entities. Unlike indexing tools such as page numbers, which are based entirely on the form of the book, the definition of entities and their relations in the *Biblical concordance* is based on concepts and their meaning. These are semantic enhancements.

A semantic *digital* publication creates the same type of enhancements, only going further using digital technology. Like the *Biblical concordance,* it doesn't merely present the

[94] M. Mueller, 'Digital Shakespeare, or towards a literary informatics', *Shakespeare*, 4:3 (2008), p. 285.

[95] S. Ramsay, 'Databases', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, Oxford and Australia: Blackwell Publishing, 2004), p. 177.

information, but intends to give the reader alternative ways of accessing and examining the text.[96] Metadata is the practical tool for this exercise in the digital world. This is not the bibliographic metadata previously discussed whose primary purpose it was to make the book as an item (rather than its text) discoverable online. In contrast with the 'general purpose metadata' of the bibliographic information, the *Book-hunter* database can be described as a 'local metadata schema': one which has a 'specific purpose … [and is] devoted to [describing] particular information objects within a very particular (local) project …'[97] While the bibliographic metadata merely provided a list of attributes of the monograph as a whole (and perhaps its individual chapters), the local metadata makes enhancements directly to the monograph's text, for example by enriching it with elements like <p topic="Thomas Dibdin">[98] which relate semantic information to the computer, in this case indicating that a specific passage is about a specific book-collector. Local metadata enhances the discoverability of entities within the digital object, rather than the discoverability of the object itself.

The semantic connections constructed with this metadata are based on so-called *subject-predicate-object* triplets, a principal component of the Semantic Web.[99] Some scientific publishers request information about these entity relations to be handed in with articles. In the realm of biochemistry, this could refer to a list of proteins mentioned in an article and their relations to one another.[100] In the case of the *Book-hunter in London*, it could mean linking a specific book to a specific book-collector. The *book-collector* is the subject; the *book* is the object; and the relationship is that the book-collector *owns* the book.

Since the results of creating a database should be structured data, there needs to be a framework for these types of entity-relationships to be fitted into. This framework is called an ontology. The ontology is a data structure designed to represent the various categories of

[96] O. Boonstra, L. Breure and P. Doorn, *Past, present and future of historical information science* (Amsterdam: NIwi-Knaw, 2004), p. 16.

[97] E. Méndez, 'Metadata Typology and Metadata Uses', in M. Sicilia (ed.), *Handbook of Metadata, Semantics and Ontologies* (Singapore and Hackenstack, N.J.: World Scientific Publishing Company, 2013), p. 20.

[98] For simplicities sake, the example given is of an XML code like the one in the introduction, rather than an RDF code which is typical of the Semantic Web. RDF can be expressed in many ways, among them through RDF/XML: 'the original standard way of representing RDF graphs'. E. Hyvönen, *Publishing and Using Cultural Heritage Linked Data on the Semantic Web* (California: Morgan & Claypool, 2012), p. 22.

[99] This is the structure of RDF, the Resource Description Framework: 'a standard model for data interchange on the Web.' (W3C, 'RDF', <https://www.w3.org/2001/sw/wiki/RDF> (1 July 2016)). The Semantic Web is a 'web of data' rather than a 'web of hypertext'. In the original hypertext form of the World Wide Web there are only links between HTML documents. On the Semantic Web it is however possible to link 'between arbitrary things described by RDF'. These things, be they objects or concepts, are represented by URIs: Uniform Resource Identifiers. T. Berners-Lee, 'Linked Data', W3, 27 July 2006, n.pag. <http://www.w3.org/DesignIssues/LinkedData.html> (1 July 2016).

[100] A. De Waard, 'From Proteins to Fairytales: Directions in Semantic Publishing', *Semantic Web*, 25:2 (2010), p. 83-84 <https://www.computer.org/csdl/mags/ex/2010/02/mex2010020083-abs.html> (28 April 2016).

information within a specific domain of knowledge.[101] In the case of *The Book-hunter in London,* authors, publishers, bookstores and book collections would for instance form separate categories in the ontology. The relationships allowed between these categories can also be defined in the ontology. The ontology can for instance state that the category *book* is the only category whose entries (for example *A Tale of Two Cities*) can form a part of a *book collection.*

In the case of the database created for this thesis, each of these categories in fact represents a table of data, into which new entries can be added. All books which appear in *The Book-hunter in London* are listed in the table 'book'. Individual entries in this table can then be linked to a table which specifies that a given book belongs to a specific 'book collection'. All the categories/tables and the links between them together form a *relational database.*[102]

How should an ontology for a historical work like the *Book-hunter in London* be formed? The historian and information scientist Manfred Thaller has argued that creating ontologies for historical sources differs in important respects from creating ontologies for scientific research. The major difference is that, unlike for instance a chemist, a historian cannot create his own data. If an initial chemical data-analysis turns out to be too imprecise, the experiment can be redefined so as to obtain the relevant missing data. The same is not possible in historical research: the knowledge which has survived from bygone times is always fragmentary. While a chemist can legitimately pre-define data slots and then fill them with data resulting from experiments, a historian using a similar method on his subject is in fact engaged in creating a new narrative, in some sense independently of the historical source.[103]

The data found in the source is out of the historian's control and much of it will not fit his pre-determined relational database. Scientists design their experiments knowing what they want from their data. Historians generally do not have a fully developed idea of what can be done with their data. Different users might use it in different ways. Thaller argues that we should therefore refrain from imposing our own interpretation upon historical data and risk

[101] Boonstra, Breure and Dorn, *Past, present and future,* p. 102.

[102] The database for the *Book-hunter in London* has been constructed as a relational database mostly for practical reasons: an RDF based data model would have taken too long to design and implement due to the technical issues involved. The relational database is modeled in a similar way to the Semantic Web, but it has a somewhat different data model. One fundamental difference is that the entities in a relational database are generally not represented by URI's (Uniform Resource Identifier), a building block of RDF. This distinction is mostly relevant for data retrieval and interoperability between different databases. The relational database has been widely used in historical research and works well for a project centred on a single publication. Both the relational database and the Semantic Web use classes and entities to code connections between digital objects: They both share the central quality of making meaning machine-readable. T. Berners-Lee, 'What the Semantic Web can represent', W3, September 1998, n.pag. <http://www.w3.org/DesignIssues/RDFnot.html> (25 June 2016). Hyvönen, *Publishing and Using Cultural Heritage*, pp. 98-100.

[103] C. Harvey and J. Press, *Databases in Historical Research. Theory, Methods and Applications* (New York: Palgrave Macmillan, 1996), pp. 190-191.

missing potentially important information in cases where it does not fit our database's ontology. Thaller calls the chemist's working practice a *model-oriented* approach: it uses a static ontology where all allowed categories and relationships are pre-defined. In the case of a history project, the *model-oriented* approach means that the encoder cannot modify his ontology during the process of encoding the source text.

Thaller believes that a *source-oriented* approach is more appropriate for historical sources, a process in which the ontology is malleable during the process of encoding the source text, and where the text can be coded many times for different purposes. The source-oriented approach aims to capture as much of the information found in the text as possible: the creator of the database doesn't predefine what will be done with it and wants to put as many tools into the researcher's hands as possible.[104]

Although there is clearly a 'tension between the complexity and the irregularity of historical sources on the one hand, and the rigid nature of the relational database on the other ...',[105] the precision of the data in a publication such as an academic monograph should be determined by its perceived usefulness for an intended audience. Keeping in mind the value of semantic editions in the humanities as outlined in chapter one, which model would be desirable for *The Book-hunter in London*? What will the average reader who consults the book's database be looking for and what can he reasonably hope to gain from it?

There are different ways to look at a publication like *The Book-hunter in London*. On the one hand, it is a secondary source. It reviews previous literature and gives an overview of its subject. If we focus on this aspect of the book, getting a precise and comprehensive coverage of its data, as a source-oriented approach would demand, seems less important than giving an overview of the network of information contained in the book, so that readers can pursue the different strands of information provided by the author.

Someone who picks up a copy of *The Book-hunter in London* is presumably looking primarily for information pertaining to a few broad categories: Information on the place London, information about books in London, and information about the people and institutions which interact with these books. A quick glance at the *Book-hunter's* cover description reveals that it contains a historical overview, so the user may be particularly interested in the state of these things at a certain point in time. A coding of *books*, *people*, *institutions* (i.e. publishers, libraries) along with some temporal and spatial framework could live

[104] Ibid. An alternate name for the *model oriented* approach is *goal oriented* approach. Meroño-Peñuela et al., 'Semantic Technologies for Historical Research: A Survey', n.pag.

[105] Boonstra, Breure and Dorn, *Past, present and future*, p. 47.

up to these needs of the user. The database would centre around, but not be limited to, these categories.[106] A model oriented approach, where entities relevant to these categories and their relationships are determined, would be perfectly serviceable to meet these ends.

On the other hand, a text from 1895 is now 120 years old and has in some sense become historically valuable in and of itself. The text gives an idea of how a man in a certain social position in late 19th century London saw the history of book-collecting, of his social assumptions, his inclusion and omissions of books and people. The author's writing becomes a way of analysing his own contemporary period. A user primarily interested in this aspect of the text would benefit from an ontology which does not make any presumptions about the categories found in the text, but one which is constructed in the process of coding it: a source-oriented approach.

Should it be the goal of enhanced editions to encode as much of the information contained in their texts as possible? As we saw in the first chapter, semantic editions should be seen as an extension of the physical book, they are a tool to make better use of the printed text. The semantic edition is not aimed at precise and/or quantitative studies of the text, although it may in some ways be useful for these purposes. It should enhance the reading experience and provide a better access to and understanding of the content. It primarily supports qualitative research.

Another reason for focusing on the qualitative side of things is the limited scope of quantitative analysis which can be performed on a text like *The Book-hunter in London*. The results of a quantitative research done on the text would mostly reflect on the author of the book, since his language and his choices determine the writing. Any kind of precise linguistic or cultural analysis of the text would give reliable knowledge of the writer himself, but could not reliably quantify anything else. More general conclusions in historical research can only be reached if 'a particular domain is completely formalized'[107] and the coverage of that domain is exhaustive. For instance, since the book *The Letters of Thomas Browne*[108] includes all of its protagonist's correspondence, it is an exhaustive corpus of knowledge. It contains all known objects relevant to this particular domain. The same cannot be said for a secondary source like

[106] These are based on the results of the 'upper-level categories' defined by researchers in a study on creating ontologies for cultural heritage. The study focused on the needs of teachers in social studies and the categories are a result of interviews with them. Aside from being in line with humanist's search queries in databases, as discussed below, this seemed a good indication of where to put the focus in an edition primarily concerned with the end-user. Their categories were nevertheless slightly different: 'time, domain concepts, people and space'. M. C. Pattuelli, 'Modeling a Domain Ontology for Cultural Heritage Resources: A User-Centered Approach', *Journal of the American Society for Information Science and Technology*, 62:2 (2011), p. 320.

[107] Meroño-Peñuela et al., 'Semantic Technologies for Historical Research: A Survey', n.pag.

[108] T. Browne and G. Keynes, *The Letters of Thomas Browne* (London: Faber & Faber, 1946).

The Book-hunter in London, where the selection of samples (such as books) depends on the preferences and access to information of a single person. As pointed out by a group of digital humanists, the role of modelling in digital publishing:

> … is not to reproduce the original in full. The only way to access the original is to visit the library, gallery, or whatever, and we should not pretend that the digital can achieve the same thing. The model's role should be to simplify, to make a complex case tractable so that we can analyse, manipulate, and communicate it more effectively and not get bogged down in irrelevant details.[109]

If we try to distinguish the core of the source-oriented approach to the book, rather than encoding everything we might come across in the text, the most important data on the cultural view of the 19th century arguably resides in the same categories that we picked out for the model-oriented approach: to know which books, people and institutions are included in the book, and at which point in time these entities appear, is to know to a great degree the focus and cultural viewpoint of the writer. A model-oriented approach seems to capture at least the essential aspects of these two ways of analysing the book. These decisions about the database are furthermore supported by research showing that in their search queries, scholars in the humanities primarily look up 'named individuals, geographical terms, chronological terms, and discipline terms'.[110] The academics can then refer to the pages where these markers are found and discover other valuable material through actual reading.

As we have seen, in the case of the monograph, semantic enhancements can be particularly helpful in a few specific ways. Among other things they help contextualize the content and to find new topics of research. These objectives remind one of the similarities between the job of a historian and a journalist. Researchers have noted that the encoded metadata in a work of cultural heritage essentially boils down to four questions, all of which count among the standard inquiries of a journalist working on a story: What, who, where and when.[111] This framework omits the *why*, which we will also get to in a later chapter.

2.2 Books, people and locations: What and who?

In a book whose subject is book-collecting, the most fundamental categories are *books* and *people*. All other coded information, such as that which has to do with *space*, *time* and *locations*, can be seen as merely a way of clarifying the context of these two categories. When we ask which questions the database should be able to answer, it therefore seems appropriate to depart from these categories.

[109] Blanke et al., 'Digital Publishing Seen from the Digital Humanities', p. 19.

[110] M.J. Bates, D.N. Wilde and S. Siegfried, 'An Analysis of Search Terminology Used by Humanities Scholars: The Getty Online Searching Project Report Number 1', *Library Quarterly*, 63:1 (1993), p. 1.

[111] M. Thaller, 'Which? What? When? On the Virtual Representation of Time', in M. Greengrass and L. Hughes (eds.), *The Virtual Representation of the Past* (UK and USA: Ashgate, 2008), p. 116.

The other advantage of these categories is their relative clarity. Classification consists of 'grouping together objects which share properties' and '[separate] objects with unlike properties into separate classes'.[112] The category *space* is very ambiguous. For instance in the case of a location of birth, it could be classified as the hospital, the city, the country of birth and so on. The category *time* could refer to the exact minute of birth, the hour, day, year or period. Although books and people can be described to a varying degree of precision, the text and design of a book is fixed after printing and a person's DNA is fixed from birth. There is a degree of ambiguity when it comes to books: a book can be described at the level of a single object, a print-roll of a single edition, an edition of a single text etc. But books do nevertheless have permanent core properties which are not as loose as those of time and space. In whichever version it is presented, the novel *Oliver Twist* is a clear point of reference. The categories books and people refer to something more tangible than space and time.

Locations mentioned in the book can also represent the 'who?', in particular institutions which are relevant to the book trade, such as publishers and libraries. This category is theoretically slightly more difficult to code in that a person remains a person, but if a private individual buys the entire catalogue of a public library, is it then still a 'library' or has it become a 'private collection'? On top of that, many publishers have also done other jobs such as bookselling and auctioning, leaving the distinction between institutions in the book unclear. Definitions aside, though institutions are not the central theme of *The Book-hunter in London*, their relation to books is nevertheless important, and likely to be of interest to end-users. Locations such as bookshops and homes of publishers and bibliophiles are therefore included in the semantic coding, and linked to the people who have a connection with these places.

The variety of questions which a search engine can answer depends on the design of the relational database. It is important that 'semantic representations ... consider *context* and *source structure* ...', since these pose limits on the questions which the database can answer.[113] In the process of what could be termed 'ontological engineering', designers apply *competency questions* to assess whether an ontology of a database fulfils its function. The database should be designed so as to be able to retrieve the answers to these questions, though they act more as guidelines for the design, rather than to limit the database to delivering only those particular

[112] C.M. Sperberg-McQueen, 'Classification and its Structures', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, UK and Australia: Blackwell Publishing, 2004), p. 161.
[113] Meroño-Peñuela et al., 'Semantic Technologies for Historical Research: A Survey', n.pag.

answers.[114] What would we for instance want to know about people, books and institutions? Here are some basic competency questions:[115]

People	• Book-collector: Which books did he or she own? • Author: Who wrote the book? Male or female?
Books	• Which books does the author of *The Book-hunter in London* mention? • Which of the plays written by William Shakespeare are mentioned in the book?
Institutions (locations)	• Bookshop: Who owned it or did business with it? • Library: Where was it located? Which books did it have? • Home: Where did a particular book-collector live?

These are very concrete questions which the database can give direct and unambiguous answers to making use of the links between different tables in a relational database. The questions can be answered directly and comprehensively by the computer because they allow for explicit answers: they can be expressed through an algorithm.[116]

In this way, the *model-oriented* approach defines which type of information will be coded and which disregarded. To answer the competency questions the database has to encode the interactions between *books*, *people* and *locations* in the relational database. This is the core of the database's content. Two other categories form a necessary backdrop to the core categories: Interactions in the world happen in the dimensions of *time* and *space*.

2.3 Time and space: When and where?

The humanities are severely disadvantaged when it comes to structuring a database, in that they generally do not have 'structured terminology and rigorous classification'. Whereas many fields of the sciences, such as engineering and biology, have sufficiently universal terminology to create shared semantic ontologies, disciplines such as history work with many terms whose meaning can be both vague and fleeting.[117] We have previously alluded to the fact that the categories of *time* and *space* are less straightforward than *books* and *people*. Time is commonly viewed as linear and one-dimensional, 'flowing in a single direction'. A computer algorithm demands that lines are drawn on explicit points of this continuum, which are not always present in historical periodisation. Space meanwhile is not just a continuum, but 'can flow in any direction', which in some cases makes the task of explicitly defining the boundaries even more

[114] Pattuelli, 'Modeling a Domain Ontology for Cultural Heritage Resources', p. 319.

[115] Categories and types of questions based on those provided in this discussion on ontologies in the humanities: V.R. Benjamins et al., 'Cultural Heritage and the Semantic Web', in C.J. Bussler, J. Davies, D. Fensel and R. Studer (eds.), *The Semantic Web: Research and Applications* (Berlin and Heidelberg: Springer, 2004), p. 435.

[116] Van der Weel, *Changing our textual minds,* p. 143.

[117] Pattuelli, 'Modeling a Domain Ontology for Cultural Heritage Resources', p. 320.

difficult than defining time.[118] These categories, which in scientific research are clear and defined by the researchers, are often ambiguous and unclear in humanities research. In historical sources, time is often not known to any degree of certainty.[119] Sometimes we only know that an event occurred in a certain interval of time. Sometimes sources differ on the timing of a single event.[120]

The assumption of our database design is that the reader will be primarily interested in books and their collectors, whether the collectors are people or institutions, and only interested in other information in as much as it contextualises these categories. The encoded temporal framework should therefore be restricted to those timings (i.e. years) which are both unambiguous and relevant to the database's main categories. For instance, the table 'people' contains the fields: 'date of birth' and 'date of death' (when known). The other chief category, *books*, has the year of issue of a book encoded if it is mentioned by Roberts.

If a user simply wants to look up a specific year, this can easily be done through full-text searching. Centuries cannot be searched in the same way, which is why temporal encoding of the century/ies in which a particular place existed, such as a bookshop, is useful for contextualisation.

The most precise unit of time encoded is a year. The book spans a long time and doesn't narrate any long linear chain of events. The situation would be different if we were dealing with a narrative in a much smaller timespan, where day-by-day tracking would be useful. However, in the case of *The Book-hunter in London*, a detail in timing more precise than a year has no obvious benefit for the end-user.

Next we come to the dimension of *space*. Ian Gregory, a specialist on the use of GIS (Geographical Information Systems) in historical research, points out the multiple uses of this system. It can be used to examine the more technical/mathematical properties of space, but it also forms the basis for 'data to be managed, visualized and analysed in ways that stress spatial relationships'.[121] This is the sense in which geospatial coding is useful in a semantic version of *The Book-hunter in London*. Visualisations of the most essential elements of a monograph help contextualise the content.

The Book-hunter in London defines its spatial limits in the title. The contribution of geospatial coding in this project is mainly to determine precise locations in the city of London

[118] I. Gregory, 'Using Geographical Information Systems to Explore Space and Time in the Humanities', in M. Greengrass and L. Hughes (eds.), *The Virtual Representation of the Past* (UK and USA: Ashgate, 2008), p. 140.
[119] Thaller, 'Which? What? When?', p. 115.
[120] Ibid., p. 117.
[121] Gregory, 'Using Geographical Information Systems', p. 139.

for one of our main categories: *locations*. The most important kind of locations for our purposes, bookshops, publishers and libraries, have the advantage of being less mobile than *books* or *people*, and it is therefore easier to assign fixed geospatial points to them.

Because they can be relatively easily mapped, these institutions benefit from an additional coding of space and time. *Space* and *time* are encoded as *attributes* to the *institution* entities. Each category has its own table in a relational database and each column on that table represents an attribute. In the case of a publisher, the attributes are 'century' and 'location'.

As previously noted, space can be described at various levels of precision. Descriptors such as *continent, country, city* and *village* are possible.[122] Despite *The Book-hunter*'s focus on London, Roberts does mention other countries and places. Space has therefore been encoded at three levels of granularity: *country, city* and *place in London*, the last one being by far the most important. The rule is coding only the categories an average reader is likely to pursue. Therefore, only the book's central subject in terms of space, London, gets a detailed coding.

2.4 Encoding rhetoric and argumentation

Finally, we get to the last of the journalist's questions: Why? The semantic enhancements we have explored so far do two things: They define entities and their relation to one another. But could semantic encoding be taken a step further? Anita de Waard, Elsevier's Disruptive Technologies Director, beliefs the problem is that these enhancements 'help us find information, but they don't help us understand it'. The point is not that we can't understand the text if we read it, but that its argumentation, the scientific discourse, isn't outlined in the coding.[123] De Waard notes that '[m]eaning is not embedded within words but rather is triggered by them', the understanding or de-coding of the information happens in the mind of the reader.[124] The genomicist Tudor Groza puts the same point succinctly, explaining that the 'discourse structures are trapped within the content of the publications, thus making the semantics discoverable only by humans'.[125]

With the ever growing output of scientific publications, Groza believes that encoding the rhetoric and argumentation of scientific papers can help researchers in picking out useful material. He wants to engage scientific writers in 'semantic authoring', which enables them to 'express their thoughts in a more structured manner', and connect their own argumentation explicitly to specific points in other argumentations, thus 'creating argumentative discourse

[122] Pattuelli, 'Modeling a Domain Ontology for Cultural Heritage Resources, p. 322.
[123] De Waard, 'From Proteins to Fairytales', p. 83.
[124] Ibid., p. 84.
[125] Groza, *Advances in Semantic Authoring and Publishing*, p. vii.

networks'. The reader benefits from a clearer understanding of the author's intended argument and better search query results.[126]

De Waard points out that if a scientific paper is viewed as an argumentative structure, it can be seen as a narrative. This narrative can be compared to a fairytale: a scientific research question acts as a protagonist; the attempts to give an answer to that question act as episodes in the fairytale; and the eventual claim made by the scientific paper is equivalent to 'the moral of the story'.[127] De Waard also notes that 'the way in which experimental scientific papers convince readers of their core claims is by using data'. The narrative and the data are dependent on each other: 'a data set by itself is not going to transmit any knowledge'.[128]

We've already seen that although quantitative research can be of use in the humanities, such as in digital humanities projects, the central arguments are usually not supported by data. In the case of a monograph, a central claim could play the role of a protagonist. A monograph's central argument stretches through many chapters, and these could be seen as representing the episodes of the fairytale. But we run into trouble with the 'moral of the story'. A humanities monograph rarely comes to a conclusion that is as decisive or easily abstracted as that of a scientific journal article.

It is not feasible to make the kind of precise and detailed breakdown of the discourse structure that De Waard recommends, since with the 'thick description'[129] arguments of the humanities it is often difficult or impossible to explicitly link a claim to specific evidence in a way that would be useful for research. A biomedical researcher can support his claim by linking to a data set, a humanist can often at best link to a summary paragraph.

However, discourse metadata in a very simple form could fulfil the function of clarifying the central claim of a chapter of a monograph and how that claim fits with the overall argument of the monograph. This might decrease the likelihood of grievous misunderstandings when academics read chapters rather than entire books. Seringhaus and Gerstein have suggested *structured digital abstracts* as a way of tying together scientific journal articles and the data that support them. These abstracts should present 'a machine-readable XML summary of pertinent facts in the article'.[130] Although the humanities are generally not concerned with establishing facts, at least in the strict scientific sense, something like a *structured digital abstract* could fulfil the role of coding *humanistic* discourse in monographs where rhetoric is particularly relevant

[126] Ibid., pp. 3-5.
[127] De Waard, 'From Proteins to Fairytales', p. 85.
[128] Ibid., p. 86.
[129] Crossick, *Monographs and open access*, p. 14.
[130] Seringhaus and Gerstein, 'Publishing Perishing?', n.pag.

(which is not the case with a book like *The Book-hunter in London*). This could be done using a simple set of pre-defined metadata entities.[131] The elements and arguments could for instance be structured like this:

Humanities discourse model

<chapter_central_claim>	A model-oriented approach is desirable for a semantic edition of *The Book-hunter in London* (1895).
<chapter_secondary_claim id="1 [2, 3, ...]">	*People* and *books* are the most essential categories of the *The Book-hunter in London* database.
<chapter_role_in_discourse>	Examines the differences and similarities between database design in the sciences and the humanities, respectively.

Through an algorithmic comparison between the keywords and the contents of these tags, a kind of humanities discourse network could be formed.

Things are considerably more complicated with regard to enhanced editions of older material. The scientific writers cited above naturally assume that the author of the enhanced paper is present to participate in the coding, given that their object of study is new scientific research. A great number of humanities publications are by writers who have long since passed on and in these cases any kind of discourse modelling is largely out of the question.

3. Weaving: Data integration on the Semantic Web

3.1 Shared ontology standards and controlled vocabulary

One of the Internet's most useful features is linking together data from a variety of sources. According to the *Research Information Network's* report, the main improvements humanities scholars would like to see with regard to their online sources is better interlinking of collections of data.[132] Disparate data impels scholars to do multiple searches on many online platforms, resulting in 'delays in research, repetitive searching, and limited ability to draw connections between sources'.[133]

Until now, we have been looking at the characteristics of *The Book-hunter in London* and defining how to make the content within the text digitally accessible, the internal linking of data. In this chapter, we come to the integration of the book into 'a larger scholarly, temporal and spatial network':[134] how the data is woven into the Semantic Web.[135]

[131] In other words, this would be a *controlled vocabulary* for discourse. It is advantageous to have a simple structure since one of the hindrances to metadata enhancements is that 'adding metadata and/or rich structure is very time consuming'. S. Woutersen-Windhouwer et al., *Enhanced Publications*, p. 31.

[132] *Reinventing research?*, p. 29.

[133] Ibid., p. 73.

[134] Crossick, *Monographs and open access*, p. 14.

There are two main tenets of data integration. Firstly, there is the structure of the data, its ontology. A group of knowledge domains, such as science laboratories in various countries, can make information retrieval more efficient by using a common ontology. Scientific publishing has put heavy emphasis on creating an efficient web of information. It is considered vital 'that the ontologies are well-formed and interface gracefully with other pre-existing ontologies ...' A scientific field of study should have a structured way of encoding metadata, which is constantly updated within the ever evolving knowledge in the area of study.[136]

In terms of ontologies, the humanities have, at least in theory, a potential for living up to the same standards as the sciences. Specific ontology languages have been built for cultural heritage in order to integrate the heterogeneous data of the humanities.[137] An example is the *CIDOC Conceptual reference model*, which is a standard ontology for cultural heritage resources. Ontologies like the CIDOC are language independent, the names of categories and attributes can vary between domains, but the conceptual model, as expressed through 'identifying codes' remains the same.[138] Ontologies are sufficiently abstract as not to cause major problems for the humanities: the categories chosen to form an ontology can be relatively unambiguous.

Things are different with regard to the second key element of online data integration, the one which is more pertinent to the *entities* which are filled into an ontology schema: controlled vocabulary. In the sciences, it is possible to create a standard set of terms which an entire area of research can use for common reference. An illustration of this can be found in the controlled vocabulary for 'Molecular Interaction' defined by the HUPO Proteomics Standards Initiative[139] or the *Gold Book*, a 'Compendium of Chemical Terminology'.[140] Meanwhile, humanities disciplines like history deal with many concepts whose meaning is not simple, clear or stable[141] and different fields and individuals are likely to use more than one word to refer to the same phenomenon.

[135] It should be noted that this subchapter is more theoretical than the previous one since it was not possible, due to both time and technical limits, to implement many of the functions discussed in this chapter into the semantic edition of *The Book-hunter in London*. See footnote 99 (p. 25) and *Appendix* (p. 66).

[136] D. Shotton, 'Adventures in Semantic Publishing', n.pag.

[137] These ontologies are expressed through *The Web Ontology Language* (OWL), which 'is a semantic markup language specially developed for publishing and sharing ontologies on the World Wide Web' and 'is developed as a vocabulary extension of RDF'. Woutersen-Windhouwer et al., *Enhanced Publications*, p. 64.

[138] Woutersen-Windhouwer et al., *Enhanced Publications*, p. 64.

[139] Proteomics Standards Initiative, 'HUPO Proteomics Standards Initiative' <http://www.psidev.info/index.php?q%20=%20node/31> (8 March 2016).

[140] IUPAC Gold Book, 'IUPAC Compendium of Chemical Terminology' <http://goldbook.iupac.org/> (8 March 2016).

[141] Pattuelli, 'Modeling a Domain Ontology for Cultural Heritage Resources', p. 320.

For this reason, controlled vocabulary is in some ways particularly vital for the humanities. Many of its disciplines deal with old primary sources which have 'alternate spellings, abbreviations, obsolete and regional word usage, idioms, misspellings, and omissions'.[142] There can even be inconsistency within the same primary source, due to the process of casting-off which became necessary with the advent of printing.[143] To avoid losing out on relevant information, and more generally for the integration of humanities publications, a common vocabulary is an important tool.

A *controlled vocabulary* is focused on a specific domain of knowledge (such as medieval history, bio-medicine or *The Book-hunter in London*) and consists of a collection of words and phrases which describe that domain. The words and phrases are tools for the indexing of texts in the domain, to facilitate access to their contents, and above all, to ensure that use of terminology is consistent across different online platforms.[144] If for instance one source speaks of *Isak Dinesen* and the other of *Karen Blixen*,[145] a search engine algorithm has no way of knowing that these names refer to one and the same person, without there being an authoritative definition of that person's identity in a *controlled vocabulary* which both publications make us of.

It is exactly this explicit definition demanded by the controlled vocabulary which creates problems for the humanities. There are in some ways less possibilities of universal consensus among humanities scholars. A pertinent example is the academics' division of history into specific periods. The periodisation of European history is neither universally applicable nor is there consensus on it among those involved in studying it.[146] However, the problem is not only a lack of consensus among historians, but perhaps even more so the fact that historical concepts have an unstable meaning. Computer scientists speak of 'concept drift' when referring to this

[142] S. Bair and S. Carlson, 'Where Keywords Fail: Using Metadata to Facilitate Humanities Scholarship', *Journal of Library Metadata*, 8:3 (2008), p. 251.

[143] After the advent of print, all pages had to be 'cast-off' before printing took place. That is, a plan had to be made for what text would go on which page. This fact, and the limited catalogue of type, meant that compositors would often create alternative spelling and abbreviations, so as not to derail the system. Hellinga, 'The Gutenberg Revolutions', p. 210.

[144] P. Harpring, *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works* (Los Angeles: Getty Research Institute, 2010), p. 12.

[145] Karen Blixen (1885-1962) was a Danish author and Isak Dinesen was her pseudonym. 'Isak Dinesen', Encyclopedia Britannica. Britannica Academic, n.pag. <http://academic.eb.com/EBchecked/topic/163827/Isak-Dinesen> (24 April 2016).

[146] J.H. Bentley, 'Cross-Cultural Interaction and Periodization in World History', *The American Historical Review*, 101:3 (1996), pp. 749-750.

evolution of concepts, it is a challenge to '[map] drifted concepts correctly ...'[147] It is harder to hit a moving target.

We have already seen the ambiguity inherent to categories such as *institutions* and *books* in the relational database. This conceptual vagueness is however particularly pertinent when concepts are not only being defined for the sake of a single publication, but for the purposes of authoritative definitions shared between many publications. And not only at the level of categories, such as *books* and *people*, but at the level of individual entries on the data tables of those categories: names of people, places, ideologies etc.

But what is the extent of this vocabulary problem of the humanities? The use of language in the humanities can be put to quantitative study. The biographer Stephen Wiberley examined the issue by looking at encyclopaedias and dictionaries which specialized in the humanities and the semantic precision of their vocabulary. Nouns turned out to be the most prominent word category. They divide into common nouns and proper nouns.[148] The most precise grammatical category is the 'singular proper term', referring to 'the name of a unique entity, either of a person or a single creative work'.[149] Wiberley found that 58% of the terms in his chosen sample corresponded to this category. This means that at least with regard to these precise terms, the coding of humanities subjects is a straightforward exercise.[150]

Research indicates that academics in the humanities have to some extent instinctively grasped this. Examination of their database usage found that search queries on 'named individuals, geographical terms, chronological terms, and discipline terms' were a lot more common among humanists than among scientists,[151] and the category of *names* seems to be by far the most important one. 86% of the singular proper terms examined by Wiberley were people's names.[152] The exactness of this category means that names can act as point of departure in humanities research:[153] if they are derived from a controlled vocabulary, they are likely to provide researchers with relevant search results.

In terms of precision, the next category below the singular proper term is the *enumerable* proper term, which refers to 'a group whose membership is so restricted that it can

[147] A. Meroño-Peñuela, 'Semantic Web for the Humanities', in P. Cimiano et al. (eds.), *The Semantic Web. Semantics and Big Data* (Berlin and Heidelberg: Springer, 2013), p. 646.
[148] Wiberley, 'Subject Access in the Humanities', pp. 420-422.
[149] Ibid., p. 423.
[150] Ibid., p. 430.
[151] Bates, Wilde, and Siegfried, 'An Analysis of Search Terminology Used by Humanities Scholars', p. 1.
[152] Wiberley, 'Subject Access in the Humanities', p. 430.
[153] J. Flanders et al., 'Names Proper and Improper: Applying the TEI to the Classification of Proper Nouns', *Computers and the humanities*, 31 (1998), p. 285.

be completely enumerated as a list of singular proper terms'.[154] An example of such a term is *The Beatles*. These were quite rare in Wiberley's sample, representing only about 1%.[155]

The problem of concept drift is mostly relegated to the least precise category of proper terms: general proper terms, which are semantically 'at the same level of precision as common terms'. The general proper term refers to words and phrases which 'designate complex and ill-defined entities'.[156] These terms are often 'collective, ideological, geographical, institutional, and cultural'. There is no way of exhaustively listing all individuals who take part in a cultural activity. *German philosophy* is not a concept whose constituent parts can be defined explicitly. In a discussion on geographical proper terms, such as *French music*, there is no set and finite list that determines which people to include so the list of singular proper terms will vary between different texts, in a way that it wouldn't in a discussion of *The Beatles*.[157]

Common terms, like common nouns, 'designate any one of a class of things or the class itself'.[158] These refer among other things to abstract ideas and 'concrete phenomena' which, despite their concreteness, have no stable meaning.[159] The word *printer* can for instance refer to both a machine and a person who does printing. Wiberley's research found that general proper terms and common terms represented 40.3% of his sample.[160] These are the terms which are least feasible for providing relevant search results.

According to a research from 1993, scientists and humanities scholars differ decisively when it comes to their use of online databases. Researchers in the humanities tend to look for particular periods in history, names of people, places and disciplines, all of which involve proper rather than common terms. Queries on 'individuals as subjects were very popular'. Searches for books and their authors were much less common.[161] Meanwhile, natural and social scientists seem to look up mostly what the researchers called 'other common terms', common terms which did not fit any of the researchers' other categories.[162] A study into search queries made by the *National Science Foundation* found '[v]irtually no terms for works as subjects, individuals as subjects, or geographical, chronological, or discipline terms'.[163] From a strictly grammatical point of view, the humanistic search query vocabulary therefore seems to be

[154] Wiberley, 'Subject Access in the Humanities', p. 423.
[155] Ibid., p. 423.
[156] Ibid., p. 424.
[157] Ibid., p. 424-425.
[158] Ibid., p. 425.
[159] Ibid., pp. 426-427.
[160] Ibid., p. 430.
[161] Bates, Wilde, and Siegfried, 'An Analysis of Search Terminology Used by Humanities Scholars', p. 31.
[162] Ibid., pp. 8 and 31.
[163] Ibid., pp. 31-32.

more precise overall. And since controlled vocabularies tend to focus on the more precise grammatical categories, they arguably have a greater importance to the humanities than to the sciences.

What follows is that with regard to the metadata in a humanities publication, both bibliographic and local, special emphasis should be put on a controlled vocabulary for names, the most important type of 'singular proper terms'. And since locations, historical periods and disciplines are also popular search terms these should also ideally be derived from a controlled vocabulary.

3.2 External linking and data fusion

Aside from an enhanced edition being semantically linked to the Web, information can also simply be hyperlinked into the publication. A scientific article can for example link every occurrence of a biological organism to its Linnaean classification on another website.[164] This type of external linking is simple and useful for contextualization. The 'book' table in the *Book-hunter* database includes a field for external links. These can refer to the texts of publications mentioned in *The Book-hunter in London*, when the source texts are freely available online. The fact that all these publications have gone out of copyright simplifies the matter. This is an obvious enhancement for exploring the subject of the book further and contextualizing its content.

The external linking is however more vulnerable than the internal linking (the relational database) because it is susceptible to *link-rot*: if links online aren't regularly updated, many of the web sites they are directed to will cease to exist or move to a different host. Research into the *Web of science* citation index found that only about half of all articles that were eleven years old at the time of the study still had functional URLs.[165] While the relational database is equally vulnerable in the sense that it too has to be present on a server to be accessible online, a publisher (not necessarily a commercial one) can easily guarantee the online presence of his own material, but he cannot guarantee that the external links will continue to function. External linking should therefore not be viewed as a central feature of the digital edition.

A more advanced form of external linking consists of bringing together data from within a publication and data from other online domains on a single web page, so-called *data*

[164] Shotton et al., 'Adventures in Semantic Publishing', n.pag.

[165] J. Hennessey and S.X. Ge, 'A cross disciplinary study of link decay and the effectiveness of mitigation techniques', *BMC Bioinformatics*, 14 (2013), n.pag. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S14-S5> (29 April 2016).

fusions.[166] In the life sciences, this type of enhancement has for instance been used for 'enhancing entities', those defined in the publication's ontology (such as *genes* or *proteins*). When the mouse passes over an enhanced entity, information is retrieved from disparate online sources and gathered into a pop-up window.[167]

The same enrichment has been provided for the core categories in *The Book-hunter in London*: *books* and *people*. This data fusion aggregates all the data about these digital objects which can be retrieved from the database into a single window, along with encyclopaedic information taken from *DBpedia*, an application which retrieves texts from Wikipedia.[168] The information that is retrieved for a data fusion has to be either licensed for use or open-source. For this project, the obvious approach was therefore to make use of the Internet's freely available material.

Conclusion

Semantic enrichment of the text of a book consists of coding its content (the relational database) and weaving the publication into the Semantic Web. Many factors call for a model-oriented approach to designing a database for *The Book-hunter in London*, an approach which uses a pre-defined ontology, and therefore sifts out the information perceived to be most relevant in a given publication. A semantic publication is geared toward a general reader, who will presumably be more interested in the book as a secondary source than he is in quantitative research of it. A decision to limit the scope of the semantic enhancements to the central categories of *books*, *people*, *locations* (particularly institutions), *space* and *time* makes possible the retrieval of data connected with the book's most important themes and maps out their context. At the same time, the database gives an idea of the scope of the writer's interests and emphasis.

The usefulness of these categories for digital indexing and information retrieval depends to a great degree on their semantic precision. From a semantic point of view, *books* and *people* are the most precise categories because they refer to concrete and relatively stable objects in the world. They constitute primarily 'singular proper terms', the most precise category of proper nouns. This unambiguity of meaning makes these categories prime candidates for search queries. A query which consists of terms from these categories is likely to return the results which the user was looking for. Due to their stable nature, these same

[166] Shotton et al., 'Adventures in Semantic Publishing', n.pag. Blanke et al., 'Digital Publishing Seen from the Digital Humanities', p. 24.

[167] De Waard, 'From Proteins to Fairytales', p. 83.

[168] DBpedia, <wiki.dbpedia.org> (22 May 2016).

categories are also useful for weaving together different knowledge domains on the Semantic Web through the use of controlled vocabularies. External linking to other web sites, and fusion of data from the database and other online domains, also serve to contextualize the book's content.

It is more difficult to form controlled vocabularies around imprecise terms and those seem to be more prominent in the humanities than in the sciences. The category of *institutions* is slightly less precise than the former two categories. Publishing houses and libraries may in some sense be concrete but they are less stable as concepts, having no physical core property which *necessarily* will remain stable, such as a printed text or a human's DNA. *Space* and *time* are the least precise categories, but they serve to contextualize the other ones.

The database that results from this can only give fully comprehensive answers to questions which can be expressed via an algorithm, which are not the kind of questions humanists tend to be most interested in. The results they give can however act as building blocks for a more general interpretation of the data contained within *The Book-hunter in London*. This will be explored further in the final chapter.

Chapter 3: Application

'*The Book-hunter in London* is put forth as a contribution to the fascinating history of book-collecting in the metropolis; it does not pretend to be a complete record of a far-reaching subject, which a dozen volumes would not exhaust'.

W. Roberts, *The Book-hunter in London*, p. xiii.

How are the semantic enhancements explored in the second chapter academically useful? How can they help academics in their research in ways that a plain printed version of the same text cannot? Since *The Book-hunter in London* is a historical text, the semantic enhancements will primarily be explored from the perspective of historical research.

1. Scholarly primitives

We have previously looked at how the humanist's research practices differ from those of the scientist on an abstract level. Now we will look at how the semantic enhancements improve the humanist's research in practice. The English professor John Unsworth has attempted to abstract the most fundamental exercises that humanities scholars perform. He denominates them as 'scholarly primitives'. These are tasks such as 'discovering', 'referring' and 'sampling', exercises which form a part of the core working practices of the humanities scholar.[169] The primitives were developed further by the DARIAH project,[170] and their definitions are useful for our purposes. A few of these primitives have a direct relevance to a semantic publication in the humanities. Following up on the conclusions of the previous chapters, the most important tasks in which digital editions can assist humanities research are in summary: originality, contextualization and subject access. Unsworth's primitives, as elaborated by the DARIAH project, can be seen as explicit definitions of how to put these into practice.

The scholarly primitive 'discovery', 'what ... [humanities scholars] do in library catalogs and library stacks',[171] can be digitally assisted through the use of technologies such as 'subject-specific' and 'specialized search'.[172] Enhancements of this type help the humanities scholar discover original material, hopefully inspiring new research projects. The primitive 'referring'

[169] J. Unsworth, 'Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?', 13 May 2000, n.pag. <http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html> (26 May 2016).

[170] S. Anderson, T. Blanke and S. Dunn, 'Methodological commons: arts and humanities e-Science fundamentals', *Philosophical Transactions of the Royal Society*, 368 (2010), p. 3789. DARIAH stands for 'Digital Research Infrastructure for the Arts and Humanities'. The project was started for the purpose of '[seeking] to understand scholarly information and knowledge practices to inform the development of a research infrastructure that is rooted in, and supports, arts and humanities research practices across Europe'. Ibid., p. 3779.

[171] Unsworth, 'Scholarly primitives', n.pag.

[172] Anderson, Blanke and Dunn, 'Methodological commons', p. 3789.

(i.e. linking) is supported by the contextualization provided by the database's semantic network. The network also provides support for the scholarly primitive of 'sampling', 'the result of selection according to a criterion'.[173] The *Book-hunter* search engine is specifically designed to make use of the enhanced discoverability of subjects within the *Book-hunter*'s text. The following discussion will be focused on how scholars could apply these scholarly primitives in their research on the *Book-hunter* database.

To test the semantic enhancements and make them openly available, a *Book-hunter* website has been created for the purposes of this project.[174] It provides the reader with the most essential functions discussed in the previous chapters: The text is presented in machine-readable form and can be browsed page by page. The opening site has a function for moving straight to a specific page number, so that a reader using a printed book can quickly access the data fusions provided for linear reading. Additionally, there is a search engine with various filters, which can be used to browse the semantic network.

It should be kept in mind, that this online edition is necessarily more limited than a professional publication would be: The digital edition has not been woven into the semantic web in any other way than through external linking. Moreover, the listing of entities which belong to categories such as *books* and *people* is not completely comprehensive. However, since the publication is aimed at qualitative research, the fact that everything is not listed down to the last item does not diminish the website's academic value, but merely limits its scope. Encoding the meaning of texts, as opposed to their form, is of course a subjective exercise which inevitably involves some manual work.[175] The work on the Book-hunter database and web site is discussed further in the *Appendix* (p. 65).

Broadly speaking, there are two main ways of using the *Book-hunter* website: exploring the text linearly or non-linearly. Looking at the text linearly refers to the more traditional way of using a book: reading it page by page and then making use of the semantic enhancements provided in the text when more information is called for. A reader making use of a printed copy of a book and then breaking away from the printed page for the enhancements is using the digital edition as an extension of the printed book. This reader is engaging directly with the text.

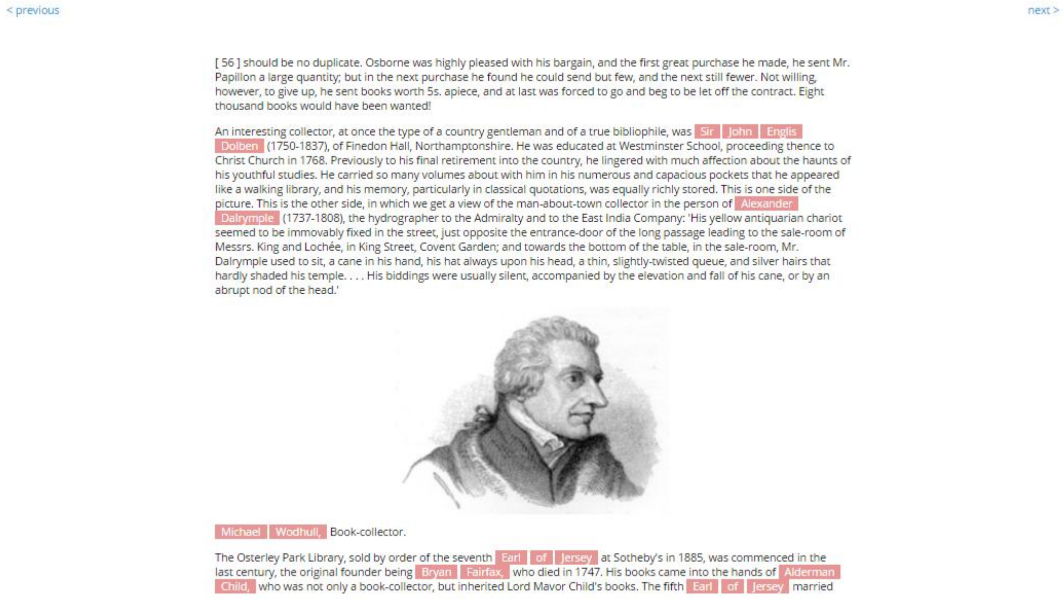
[173] Unsworth, 'Scholarly primitives', n.pag.

[174] Enhanced edition of "The Book-hunter in London", <http://bookandbyte.org/bookhunter/> (31 May 2016).

[175] Van der Weel, *Changing our Textual Minds*, pp. 157-158. In a commercial setting, substantial use is made of automatic detection for semantic enrichment. An example of a cultural heritage project which makes very extensive use of automatic extraction is *Europeana*, a common cultural heritage portal for Europe. See *Report on Enrichment and Evaluation* (Den Haag, Netherlands: Europeana, 2015) <http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf> (2 July 2016).

The other type of reader engages *indirectly* with the text. Non-linear reading refers to using a search engine to retrieve specific information from the text. Many readers will of course take both of these approaches, but different tools are required for using the book in these distinct ways and they do to some extent serve different purposes in research. A separate subchapter is therefore devoted to each.

A common denominator among the tools provided on the *Book-hunter* website is that they are not doing any truly innovative tasks. They are performing tasks which in theory could be done manually, but doing them more efficiently and economically. Studies suggest that academics in the humanities 'adopt a technology when it improves upon their current practice'. They are likely to engage with a product that improves on what they are already doing more so than one which provides completely new research possibilities.[176] We will therefore focus on how the semantic edition improves upon traditional research methods, rather than how digital technology can be innovative in research.[177]



[56] should be no duplicate. Osborne was highly pleased with his bargain, and the first great purchase he made, he sent Mr. Papillon a large quantity; but in the next purchase he found he could send but few, and the next still fewer. Not willing, however, to give up, he sent books worth 5s. apiece, and at last was forced to go and beg to be let off the contract. Eight thousand books would have been wanted!

An interesting collector, at once the type of a country gentleman and of a true bibliophile, was Sir John Englis Dolben (1750-1837), of Finedon Hall, Northamptonshire. He was educated at Westminster School, proceeding thence to Christ Church in 1768. Previously to his final retirement into the country, he lingered with much affection about the haunts of his youthful studies. He carried so many volumes about with him in his numerous and capacious pockets that he appeared like a walking library, and his memory, particularly in classical quotations, was equally richly stored. This is one side of the picture. This is the other side, in which we get a view of the man-about-town collector in the person of Alexander Dalrymple (1737-1808), the hydrographer to the Admiralty and to the East India Company: 'His yellow antiquarian chariot seemed to be immovably fixed in the street, just opposite the entrance-door of the long passage leading to the sale-room of Messrs. King and Lochée, in King Street, Covent Garden; and towards the bottom of the table, in the sale-room, Mr. Dalrymple used to sit, a cane in his hand, his hat always upon his head, a thin, slightly-twisted queue, and silver hairs that hardly shaded his temple. . . . His biddings were usually silent, accompanied by the elevation and fall of his cane, or by an abrupt nod of the head.'

Michael Wodhull, Book-collector.

The Osterley Park Library, sold by order of the seventh Earl of Jersey at Sotheby's in 1885, was commenced in the last century, the original founder being Bryan Fairfax, who died in 1747. His books came into the hands of Alderman Child, who was not only a book-collector, but inherited Lord Mavor Child's books. The fifth Earl of Jersey married

Figure 1. A screenshot from the *Book-hunter* web site. Page 56 of the book with semantically enhanced entities highlighted. Enhanced edition of *The Book-hunter in London*, <http://bookandbyte.org/bookhunter/showTei.php?page=56> (17 July 2016).

[176] *Reinventing research?*, pp. 12-13.

[177] Databases can of course be used in innovative ways for humanities research. Paul Ell and Lorna Hughes complain about a 'lack of substantive methodological innovation', 'that major advances in the digital humanities tend to relate to addressing current historiographical questions, rather than asking new questions.' The complaint is legitimate but innovation is the task of researchers rather than publishers. A published work is meant to be used by many people for many purposes, so its coding is necessarily more general than that of a research project. P. Ells and L. Hughes, 'E-infrastructure in the Humanities', *International Journal of Humanities and Arts Computing*, 7:1-2 (2013), p. 31.

2. Looking in from the outside: The search engine

2.1 Sampling the text

When readers make use of the *Book-hunter* search engine, they want specific information from the book which fits a certain criteria. This is what Unsworth calls 'sampling'[178] and is inherently different from a reader engaging directly with the text. Instead of the reader processing the text herself, she sends in a digital messenger which processes the text on her behalf. The reader is looking into the text from the outside. The search engine is the tool with which the user samples the text.

There are some reasons to doubt whether semantic enrichment of texts is actually desirable. An advanced search engine certainly improves the accessibility of a book's content, but does it also change the way in which the publication is used in undesirable ways? Research has shown that at least within the sciences, digitisation has led to '"bouncing" behaviour'. Rather than reading through an entire article, readers tend to move swiftly from one article to another through hyperlinks.[179] Further indication that people are using online resources chiefly for non-linear reading is given by *Springer*'s 2008 survey on e-book users, which found 'that the type of eBooks most frequently used [were] reference works and textbooks'.[180]

Full-text linear reading may be the ideal but what is the reality of usage? The librarian and writer Rick Anderson makes a salient point. Namely, that to an extent, monographs have always been used in this way. No matter the intention of the authors, Anderson notes that when researching a topic as a student himself, he and other students rarely read through an entire book. Instead they performed what essentially amounts to text-mining on a large roster of publications, skimming the pages for relevant information. The students back then were limited in their task by the level of detail in the index of the books they were browsing.[181] Since this is clearly the use that students make of most monographs, and also what more senior members of universities presumably do as well, the digitisation and semantic enhancements do nothing but aid the researchers, give them better tools to work with.

One of the roots of the academic culture of skimming is simply a lack of time. Before embarking on any research task, a time-constricted academic or student will do a type of cost

[178] Unsworth, 'Scholarly primitives', n.pag.

[179] *Collaborative yet independent*, pp. 16-17.

[180] 'eBooks – The End User Perspective', Springer, 2008 <http://www.springer.com/cda/content/document/cda_downloaddocument/eBooks+-+the+End+User+Experience?SGWID=0-0-45-608298-0> (15 March 2016).

[181] R. Anderson, 'Monographs as Essays, Monographs as Databases: Or, the Irrelevance of Authorial Intent', n.pag. [under 'Unpublished secondary sources' in bibliography].

benefit analysis: Will this activity be fruitful enough for the research project to merit the time invested in it?[182]

At the most basic level, the *Book-hunter* search engine is simply a hyperlinked index. Instead of having to flip to the back of the printed book to look up a subject and then flip to every page to make notes, this digital index can instantly retrieve all the references to a subject. This is certainly valuable as a time saver. However, the search engine is in this instance performing a task which can be performed *relatively* quickly with the printed version. Using an analogue index takes longer, but not so much longer that the reader will be deterred from doing it. This simple form of digital subject access only marginally lowers the threshold for this task to be undertaken.

The search engine is most useful when it dramatically lowers the threshold for an academic's decision to perform a research task, a task which in theory could be done manually but which a researcher would likely be deterred from doing due to the time and effort required. As Martin Mueller explains, 'digitization changes the time calculus of many activities'. By saving the researcher a substantial amount of manual work, the search engine increases the likelihood that a researcher's cost benefit analysis will give a positive result.[183]

This is applicable to the search engine's more advanced functions, those that perform research tasks which would have broken down any cost benefit analysis if they had to be done manually. These are functions which, unlike the index, a printed book does not even provide in analogue form. Aside from cost-benefit issues, the other condition for use is of course that the search engine can retrieve academically relevant results. Keeping both of these factors in mind, we will look at a few examples of these types of tasks performed on the *Book-hunter* search engine.

The relational database of *The Book-hunter in London* provides opportunities for filtering the search engine's results. Filtering the results through a criterion of the user's choice is the basis for examining the text in ways not possible with a printed book (except of course with a great deal of manual work).

The semantic connections in the relational database allow the search engine to go further than aggregating references to the entities found in traditional printed indexes. These indexes generally do not aggregate information about a subject beyond a certain level of abstraction. The *Book-hunter*'s printed index includes references to various women (though they are a tiny minority compared to the number of men mentioned). There are entries for individual women,

[182] Mueller, 'Digital Shakespeare', p. 287.
[183] Ibid.

and there are entries for 'Women as book-collectors' and 'Women as book-thieves'.[184] But the category 'women' does not exist, simply because subjects at that level of abstraction tend to have too many entries to be feasible in print. Not being bound by the size of a page, the digital index does however include this option. Gender is one of the attributes tin the table 'person' in the relational database. One *could* of course go through the index manually, find all the women and look up all the pages, but the speed of retrieval provided by the search engine significantly lowers the threshold for doing this research task.

The search engine can also outdo the printed index in terms of search criteria. In some cases, even going manually through the index might not get you the results you want. In contrast with the category 'women', in order to find for instance all the printers mentioned in *The Book-hunter in London,* you would have to know the names of all the major English printers beforehand, as the occupation of people is not provided in the index. Without the digital index, a non-expert on the history of English print would in all likelihood simply have to read the book to retrieve this information, but the 'printer' filter in the table 'occupations' in the relational database makes it possible to retrieve this data instantly.

A printed index provides at best a very basic hierarchy of subjects, such as listing *A Tale of Two Cities* under the heading of 'Dickens, Charles'. In contrast, the *Book-hunter* search engine allows for multifaceted searches: It is possible to use more than one criterion to filter the results. One could for instance make a query about all books written by writers who were born in the 17th century in a specific chapter of the book. The search engine is therefore particularly useful for someone who is looking for information on a very specific topic and one which is not usually listed in traditional indexes: A topic at a certain level of abstraction, such as '17th century writers', which the user believes is probably covered in the book, but is not treated specifically in any single section of it.

In these cases, the semantic enhancements clearly support the scholarly primitive 'discovery' through the enhanced *subject access.* The explicit answer the search engine gives to questions comparable to the competency questions discussed earlier, such as 'How many of the writers mentioned in *The Book-hunter in London*' were born in the 17th century?', can form a part of larger research projects, for instance a project aimed at providing an overview of arts and culture in the 17th century. In a large scale project for which a lot of material has to be explored and to which the information in *The Book-hunter in London* may be relevant but is perceived to be peripheral, the book would not be used were it not for the enhanced subject access.

[184] Roberts, *The Book-hunter in London,* p. 333.

2.2 Mapping the content

The Book-hunter in London is a historical work and particular types of digital enhancements are especially pertinent for that discipline: those that contextualize the data. To a historian, facts have no value in isolation. Studying a subject means familiarizing oneself with all its various aspects and making connections between them. What makes this process laborious is that the working memory of human beings can only juggle so many ideas at once; psychologists have estimated about three to four.[185] This means that the historian has to take notes, draw up mind-maps and chronological tables, in order to be able to see the context and make the connections.

The computer is however not as constricted in terms of its memory. Every datum that has been marked-up and located in a database can be retrieved instantly. But the computer does not understand its own enterprise and can only follow explicit orders. It can't make any of its own inferences. The historian and the computer can therefore make up for each other's limitations. The historian has a limited memory, but he has the intelligence to make the connections. The computer remembers everything, but understands nothing.

One of the tasks where the computer's memory can make up for the academic's limited recall is in visualisations. A visual display can potentially give insight into 'the complex, often nonlinear relationships among the topics, events, people and places buried within the sources'.[186] The result lists delivered by the *Book-hunter* search engine might enable us to make certain connections, by exploring the data fusions which appear when each digital object is clicked on, but the search results in and of themselves do not provide much context. In visual mapping, context is inherent to the visualisation itself: the spatial and temporal context.

Let us begin with the spatial dimension. The visualisation of certain encoded entities using GIS (Geographical Information System), for instance the category of locations called 'Book shop/Publisher/Auction house',[187] provides two main advantages for research: the visualisations allow us to examine the relations between the entities in space, and they allow us to make inquiries where locations are relevant.[188]

We might for instance be interested in information on bookshops in a particular neighbourhood of London. The GIS map dramatically lowers the threshold for exploring this topic through the use of *The Book-hunter in London.* Without the GIS mapping, the spatial

[185] S. Pinker, *The Sense of Style. The Thinking Person's Guide to Writing in the 21st Century* (UK and elsewhere: Penguin Books, 2015), p. 67.

[186] T. Lindquist et al., 'Using Linked Open Data to Enhance Subject Access in Online Primary Sources', *Cataloging & Classification Quarterly*, 51 (2013), p. 916.

[187] These three operations often co-occur in a single place and the distinction between them is often blurry. They have therefore been united into one category. This is by far the most prominent category of locations.

[188] I.N. Gregory and P.S. Ell, *Historical GIS. Technologies, Methodologies and Scholarship* (Cambridge, New York and elsewhere: Cambridge University Press, 2007), p. 18.

relations between the bookshops could only be uncovered by a combination of close reading and manual mapping.[189] The locations encoded in the relational database are also temporal: there is an attribute in their data table which defines which century/ies the given entity existed in.

This digital mapping can be useful for providing larger context in a research project which is not necessarily centred on the book trade. Since history is all about context, historians exploring events that happen in a certain area in a certain period will want to know many different facets of life in that period. We can for example imagine that we were writing a biography of a man of letters who lived in the Strand, a street in London, during the 19th century.[190] Even though the book trade were not the central issue of our research, we might want to know which bookshops were in the neighbourhood, find out who frequented them and sample which kind of books were available. This would give us a sense of the intellectual milieu in which the protagonist lived. The combination of geo-spatial mapping and data fusions (discussed further below) makes this possible.

Furthermore, geospatial mapping can uncover connections between entities which are not apparent in the source text. A good example of a history project which used GIS to this effect is Benjamin C. Ray's digital library covering the Salem Witch Trials of Massachusetts in 1692. From a large selection of digitised primary sources, Ray created a database which '[linked] ... every document, every image, and every piece of demographic and genealogical information to every person involved with their location in space and time.'[191] This enabled him and others to find patterns in the data which were implicit in the sources, but were only made clear by the digital mapping. For instance, a coding of the tax rates of the different people in the village and the display of this data on a GIS map revealed a relatively even distribution of wealth in the community, providing statistical support for the theory that social, rather than economic, factors were at the heart of the trials.[192]

The *Book-hunter* database is more limited in the sense that it represents only a single text, so it cannot make a precise comparison between different sources, and its conclusions

[189] That is of course the way in which these spatial relations were added into the database but the time of the encoder of the semantic edition is spent much more economically than that of the individual researcher, since each location encoded for the semantic edition can be used for a variety of purposes, not just to serve a single research question. The process can also be automated to a great degree. The semantic connections between the *Book-hunter* database and GoogleMaps were created automatically through the use of the controlled vocabulary GeoNames, <http://www.geonames.org/> (26 May 2016).

[190] The street still exists, but using a modern map can of course be misleading when dealing with historical data, since the urban landscape may have changed over time.

[191] B.C. Ray, 'Teaching the Salem Witch Trials', in A.K. Knowles (ed.), *Past Time, Past Place: GIS for History* (California: ESRI Press, 2002), p. 21.

[192] Ibid., pp. 22 and 25.

about historical phenomena cannot be generalised because it sees the world through the subjective eyes of a single individual.

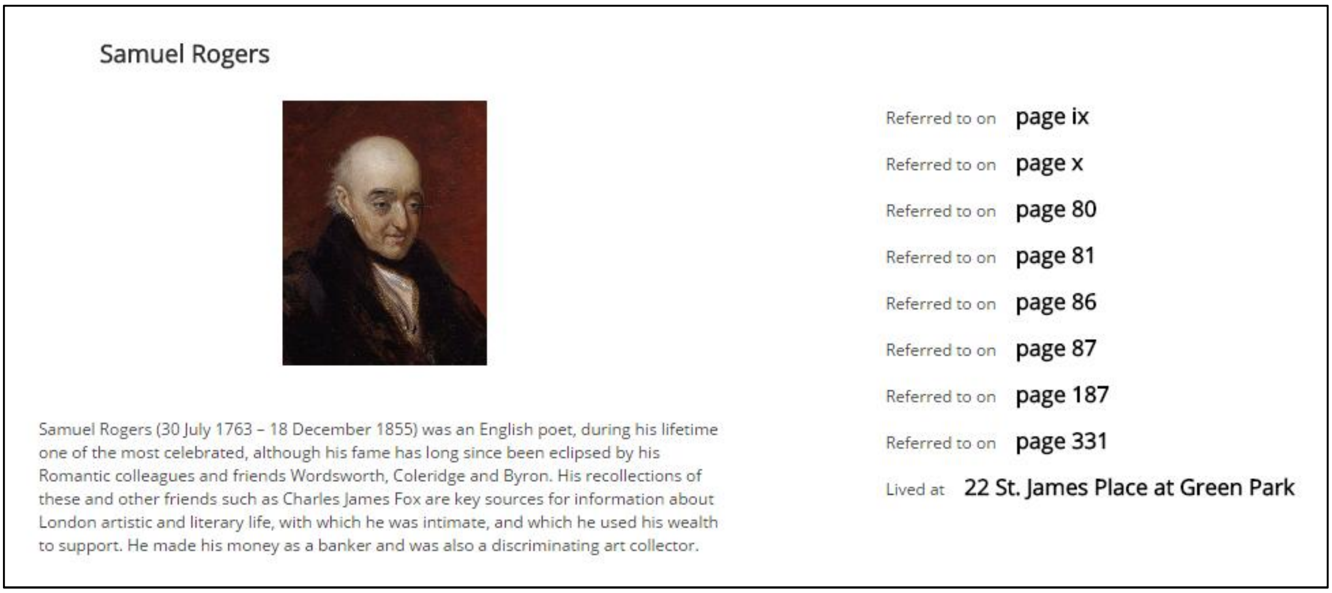
The database can nevertheless fulfil a similar function, but acting as a roadmap rather than a final destination. We know for instance, that William Roberts was very knowledgeable about the history of the book trade, having written several books on it, so we can take the bookshops he mentions in each century as an indication of how the centres for book-collecting shifted between neighbourhoods through the centuries. This is not an ultimate prove of the evolution of the book trade, but the mapping can give the researcher an idea of *where* to look. And knowing what to look for is one of the keys to a successful use of historical sources.

3. Looking out from the inside: The data fusions

The heading of this chapter refers not only to the external linking provided by the semantic publication, but also to looking out from any particular vantage point within the semantic network of *The Book-hunter in London*. Imagine that our database were a village built on a high hilltop with a good view of the surrounding area. An individual standing inside the village does not only get a good view of the surroundings, he also sees the village itself. Depending on his position within the village, he will see different streets and alleyways and his view on the surroundings will also be affected by his current vantage point. This is the sense in which a reader of the text 'looks out'.

As discussed previously, the *Book-hunter* database is a representation of the information contained in the original text. This means that the semantic edition is not a source in and of itself. Any interpretation will ultimately have to come from reading the text of the book or the contextual information. Rather, the semantic network facilitates discovery of the content of the publication's text and of related information. It reveals connections implicit in the text, but ones which the reader may not have noticed if he relied solely on the printed book. In this way, it can help the researcher to discover new dimensions to her subject, give her better context, and even lead her into original research of a theme encountered through the semantic links. Using Unsworth's terminology, the enhancements as a whole support the scholarly primitive of 'referring': linking to relevant material.[193]

[193] Unsworth, 'Scholarly primitives', n.pag.



Samuel Rogers

Samuel Rogers (30 July 1763 – 18 December 1855) was an English poet, during his lifetime one of the most celebrated, although his fame has long since been eclipsed by his Romantic colleagues and friends Wordsworth, Coleridge and Byron. His recollections of these and other friends such as Charles James Fox are key sources for information about London artistic and literary life, with which he was intimate, and which he used his wealth to support. He made his money as a banker and was also a discriminating art collector.

Referred to on page ix
Referred to on page x
Referred to on page 80
Referred to on page 81
Referred to on page 86
Referred to on page 87
Referred to on page 187
Referred to on page 331
Lived at 22 St. James Place at Green Park

Figure 2: A screenshot from the *Book-hunter* web site. Information from DBpedia and the relational database combined. Enhanced edition of *The Book-hunter in London*, <http://bookandbyte.org/bookhunter/showDataPerson.php?person=669> (13 July 2016).

The data fusions for *books* and *people* are the most important enhancement for contextualization. Their function is that a user can click on these entities where they are found in the text and a window will open, containing bibliographic information from *DBpedia* along with all the contextual information provided by the relational database.

Despite the fact that researchers making use of the online edition can quite quickly and efficiently look up books and people on other platforms, the added value of this instant retrieval of encyclopaedic information is not negligible. As a researcher, you never know where you are going to find your next lead and clicking on a list of publishers contained in *The Book-hunter in London* is substantially faster than looking them all up independently on another platform. The *DBpedia* fusion does not provide any innovative functions, but simply makes the task of retrieving contextual information more inviting, even when the researcher has only a slight suspicion that examining the context will be useful. The external links to original texts found online fulfil a similar function: they are immediately available and therefore the texts they refer to are more likely to be browsed than otherwise.

A more substantial advantage over the printed page is provided by the semantic connections retrieved from *within* the relational database: The relations of *people* with *books* and *places*, and the relation of *books* with *people* and *places*. When a user clicks on a person's name she gets an immediate overview of all the encoded information about that person which is found in the book, including semantic connections with other entities, such as that a particular person is mentioned in connection with a particular place.

The village metaphor is most pertinent to these enhancements. The principal contribution of these semantic links is to enable the user to 'look out' from the point of view of any single individual in the village, any entity in the semantic network, and see its relation to the others. If we would for instance be working on an essay about the *Mazarin Bible*, its data fusion derived from the *Book-hunter* database will give us a list of people who owned the book and refer us to the pages in the book where these people are mentioned, and we might then want to look them up individually. The data fusions can be seen as a source of inspiration, a way of browsing the semantic network without a specific goal, but simply to inspire original research projects.

Conclusion

The main benefit of the semantic enhancements of *The Book-hunter in London* is lowering the threshold for performing various research tasks, rather than enabling new types of research. This emphasis is appropriate in an edition which is intended to be used by a general audience for multiple purposes, rather than for a specialised research project where bespoke enhancements may be ideal.

The Book-hunter enhancements are not a source of information in and of themselves, unlike for instance a word frequency list generated by an algorithm. The data on the frequency of words in a text is information which can be referred to directly and used to argue a case because the information cannot be deduced from reading the text (reading is a different exercise from counting).

The semantic network of *The Book-hunter in London* maps out semantic context and connections which *can* be deduced from reading the text, but which would require extensive manual work due to the limited memory of human beings. The researcher will always have to refer to the sources to which the semantic network connects him, rather than the results which the data fusion or the search engine present him with. Ultimately, any interpretation comes from looking at the text itself.

'Discovery' is therefore the most important of Unsworth's scholarly primitives in the context of a semantic publication. The semantic connections enable the user to 'sample' the text, using search criteria which are wider and more detailed than what the average printed index allows for. The links in the relational database contextualise the entities contained within it (such as books and people), and this context can be used to provide the user with both data fusions and digital maps, speeding up bibliographic research and saving the user a great deal of manual effort. The semantic edition is a tool for exploring the text of *The Book-hunter in London* faster, more thoroughly and in better context than would otherwise be possible.

Final conclusion

The most fundamental purpose of an academic text in any form, whether it is presented as a printed book, an online edition or a pamphlet, is to inform the reader. Any technology which improves this information retrieval, clarifies what is contained in the text and provides relevant context, is furthering this goal.[194]

There are at least three potential barriers to the enhancements being put to productive use. Firstly, the potential readers of a publication have to be able to find the book and book-hunting these days happens mostly online. Books can of course be discovered in various ways, both formal and informal, but the most basic and essential tool a publisher of a given text has in his hands to increase the likelihood that the publication will find its audience is bibliographic metadata, the data which helps search engines to find relevant texts.

The online integration of the publication can be taken further than that. Both the local metadata, that which is coded directly into the text of the publication, and the bibliographic metadata can be woven into the larger network of online information principally in two ways. Firstly, this can be done through the use of a standard ontology (the data structure), so that similar semantic connections will be created between similar digital objects in different publications. Secondly, the publication's discoverability can be enhanced with controlled vocabulary, so that the names of digital objects which refer to a single entity (such as *Oliver Twist,* William Roberts or post-modernism) are written in the same way in the databases of different publications. These types of data integration are put in place at the service of multiple-purpose search engines which require some uniformity in data structure to be able to give comprehensive results on a search query.

Secondly, there is the barrier of screen-reading: most people still prefer to read monographs on paper. A uniquely digital edition of a book may well be used in a non-linear fashion: to sift out information under a certain criteria through full-text searching or a bespoke search engine. But sustained linear reading requires that there be a printed edition as well. Detailed metadata and a combined paper/e-book edition are therefore ideal for a semantic publication. Since the printed version forms a part of the equation, the semantic edition should be focused on tasks in which it can improve upon its printed counterpart.

The third potential barrier, and the most important one, is the actual content of the semantic publication: are the semantic enhancements suited to the needs and interests of the

[194] At the very least, it is furthering this goal in the theoretical sense that researchers will have better tools at their disposal to do their work. Whether the enhancements will also lead to undesirable reading habits and other negative consequences is another matter which has to be treated separately.

readers? Semantic enhancements have a more extensive history in scientific publishing than in humanities publishing, so it seems natural to make a comparison between these two broad areas of research.

A humanist reader is different from a science reader and the semantic enhancements should reflect the differences. Though the new discipline of the digital humanities may benefit from the possibilities of including interactive data in online publications, a great majority of writing in the humanities is not data driven, in sharp contrast to the sciences. The qualitative nature of humanities research means that rather than data, *context* is the central concern in a semantic humanities publication. It also means that encoding rhetoric and argumentation is more difficult in the humanities: the arguments are less explicit and most often related entirely in words rather than with data.

Furthermore, research projects in the sciences tend to be collaborative, due to the universal nature of their objects of study, while humanist studies tend to be individual, due to the culture- and context-dependent nature of their preoccupations. In comparison with the sciences, building on other people's research is less important than finding an unexplored area of interest, which means the semantic enhancements of a humanities publication have the additional objective of helping the reader to find original research ideas.

The tool which lies at the heart of the design of the *Book-hunter* enhancements is a relational database. The ontology of this database, the way in which the data is structured, has to be designed so as to be able to retrieve the contextual information which the creators of the database believe the intended audience will find useful.

Audiences are different. If the intended audience is for instance a group of historical researchers, keen to explore the world-view of William Roberts, the writer of *The Book-hunter in London*, in precise quantitative terms, then a detailed encoding of as much information as possible is ideal. A general audience is more diverse and will want to use the semantic edition for more varied purposes. For a semantic edition of a text like *The Book-hunter in London* it therefore seems appropriate to limit the semantic coding to the categories most directly relevant to the book's central theme: *books*, *people* and *institutions*, with an additional coding of the necessary contextual devices *time* and *space*.

The exact way in which these categories are represented in the relational database is similarly determined by assumptions about the publication's audience. The ontology of the database, which consists of categories (tables in the case of the relational database) and the connections between them, determines which types of questions a search engine can give explicit answers to. For instance, a table called 'profession', linked to each individual entry in the table 'people', provides the possibility for asking questions such as: 'Which *booksellers* of the 18th century are mentioned in the book?' These answers can then contribute to answering more interpretive questions.

The semantic enhancements aim to help the readers make discoveries which they would not have made, or would have been unlikely to make, with only a printed book to consult. Through the search engine which the *Book-hunter* website provides on its opening page, a researcher can make use of the enhanced *subject access* provided by the semantic network. This network is essentially an advanced hyperlinked index. It allows the user to 'sample' the text, retrieve information from it under precise criteria.

The relational database is also a supplier of context. This applies in particular to the GIS visualisations, which make use of the spatial and temporal information encoded about locations in the database. The visualisations are an alternative way of viewing the information contained in the text, and one which makes explicit context which is implicit in the pages of *The Book-hunter in London*.

Though they are a more straightforward device, the data fusions about *books* and *people* immensely increase the speed and convenience of orientating oneself in the historical context of the text. They make use of the relational database's ability to aggregate disparate data on a single subject, from outside sources (*DBPedia*) and from the inside network (the relational database), illustrating internal connections in the text (such as the information that a certain individual owned a certain book).

Both of these enhancements are an extension of the original text which are meant to orientate the reader in the book's network of information; they should help him discover and explore original ideas through contextual information and connections which are already implicit in *The Book-hunter in London*, but which elude most readers when they examine the text linearly.

It is precisely the non-linear nature of the semantic enhancements which make them a useful tool. Whether they are used during traditional linear reading of the book or not, the enhanced subject access (made use of through the search engine), the visualisations and the data fusions all share the *raison d' être* of bringing together data dotted throughout the book and from other publications into a single space. For this reason, books which rely strongly on a linear narrative are less good candidates for semantic editions than ones which present an overview of a subject where each part is to an extent self-contained.

When a text is semantically enriched for a general audience, the perceived end-user has to determine every step of the process: which text to enrich, how to enrich it and how to present the enrichments. The central aim is to provide the reader with tools which he or she can use to explore the text more deeply and efficiently, to interactively map out the information contained in the source and thereby improve the reader's understanding of the book's subject.

Ultimately, the information resides not in the semantic enhancements, which are merely a digital extension, but in the original text and the interpretation takes place in the mind of the reader. The semantic edition is there to enhance the readers' experience of the text, enabling them to examine the information contained in it more precisely than would be possible in a printed book, and to get a clearer view of the larger context, which is so central to research in the humanities.

Bibliography

Case study

Roberts, W., *The Book-hunter in London. Historical and other Studies of Collectors and Collecting* (Elliot Stock: London, 1895) <http://www.gutenberg.org/ebooks/22607> (12 April 2016).

Case study web site

Enhanced edition of "The Book-hunter in London", <http://bookandbyte.org/bookhunter/> (10 July 2016).

Published secondary sources

Allen, B., 'Paul Mellon and Scholarship in the History of British Art', in *Paul Mellon's Legacy. A Passion for British Art. Masterpieces from the Yale Center for British Art* (New Haven and London: Yale University Press, 2007), pp. 43-56.

Anderson, S., T. Blanke and S. Dunn, 'Methodological commons: arts and humanities e-Science fundamentals', *Philosophical Transactions of the Royal Society*, 368 (2010), pp. 3779-3796.

Bair, S., and S. Carlson, 'Where Keywords Fail: Using Metadata to Facilitate Humanities Scholarship', *Journal of Library Metadata*, 8:3 (2008), pp. 249-262.

Baron, N.S., *Words Onscreen. The Fate of Reading in a Digital World* (Oxford and elsewhere: Oxford University Press, 2015).

Bates, M.J., D.N. Wilde and S. Siegfried, 'An Analysis of Search Terminology Used by Humanities Scholars: The Getty Online Searching Project Report Number 1', *Library Quarterly*, 63:1 (1993), pp. 1-38.

Bekhuis, T., 'Keywords, discoverability, and impact', *Journal of the Medical Library Association*, 103:2 (2015), pp. 119-120.

Benjamins, V.R., et al., 'Cultural Heritage and the Semantic Web', in C.J. Bussler, J. Davies, D. Fensel and R. Studer (eds.), *The Semantic Web: Research and Applications* (Berlin and Heidelberg: Springer, 2004), pp. 433-444.

Bentley, J.H., 'Cross-Cultural Interaction and Periodization in World History', *The American Historical Review*, 101:3 (1996), pp. 749-770.

Berners-Lee, T., 'Linked Data', W3, 27 July 2006, n.pag. <http://www.w3.org/DesignIssues/LinkedData.html> (1 July 2016).

Berners-Lee, T., 'What the Semantic Web can represent', W3, September 1998, n.pag. <http://www.w3.org/DesignIssues/RDFnot.html> (25 June 2016).

Berners-Lee, T. and J. Handler, 'Publishing on the semantic web', *Nature,* 410:26 (2001), pp. 1023-1024.

Bhaskar, M., *The Content Machine. Towards a Theory of Publishing from the Printing Press to the Digital Network* (UK and USA: Anthem Press, 2013).

Black, M.H., 'The Printed Bible', in B.M. Metzger and M.D. Coogan (eds.), *The Oxford Companion to the Bible* (New York and Oxford: Oxford University Press, 1993), pp. 611-615.

Blanke, T., et al., 'Digital Publishing Seen from the Digital Humanities', *Logos,* 25:2 (2014), pp. 16-27.

Bluestone, M., 'AAP StatShot: Publisher Net Revenue from Book Sales Declines 4.1% in First Half of 2015', 8 October 2015, n.pag. <http://publishers.org/news/aap-statshot-publisher-net-revenue-book-sales-declines-41-first-half-2015> (19 March 2016).

Boonstra, O., L. Breure and P. Doorn, *Past, present and future of historical information science* (Amsterdam: NIwi-Knaw, 2004).

Browne, T. and G. Keynes, *The Letters of Thomas Browne* (London: Faber & Faber, 1946).

Collaborative yet independent: Information practices in the physical sciences (UK: The Research Information Network, 2011) <http://www.rin.ac.uk/system/files/attachments/Phys_Sci_case_study_full_report.pdf> (8 June 2016).

Crane, G. et al., 'Beyond the Digital Incunabula: Modeling the Next Generation of Digital Libraries', in J. Gonzalo et al. (eds.), *Research and Advanced Technology for Digital Libraries,* vol. 4172 (Berlin and Heidelberg: Springer, 2006), pp. 353-366.

Crossick, G., *Monographs and open access. A report to HEFCE* (London: HEFCE, 2015) <http://www.hefce.ac.uk/media/hefce/content/pubs/indirreports/2015/Monographs,and,open,access/2014_monographs.pdf> (11 March 2016).

De Waard, A., 'From Proteins to Fairytales: Directions in Semantic Publishing', *Semantic Web,* 25:2 (2010), pp. 83-88 <https://www.computer.org/csdl/mags/ex/2010/02/mex2010020083-abs.html> (28 April 2016).

De Waard, A., 'The future of the journal? Integrating research data with scientific discourse', *Logos,* 21:1 (2010), pp. 7-11.

Dictionary of Nineteenth-Century Journalism in Great Britain and Ireland (Gent and London: Academia Press and the British Library, 2009).

East, J.W., 'Subject Retrieval from Full-Text Databases in the Humanities', *Libraries and the Academy,* 7:2 (2007), pp. 227-241.

East, J.W., 'Subject retrieval of scholarly monographs via electronic databases', *Journal of Documentation,* 62:5 (2006), pp. 597-605.

'eBooks – The End User Perspective', Springer, 2008 <http://www.springer.com/cda/content/document/cda_downloaddocument/eBooks+-+the+End+User+Experience?SGWID=0-0-45-608298-0> (15 March 2016).

Ells, P., and L. Hughes, 'E-infrastructure in the Humanities', *International Journal of Humanities and Arts Computing*, 7:1-2 (2013), pp. 24-40.

Flanders, J., et al., 'Names Proper and Improper: Applying the TEI to the Classification of Proper Nouns', *Computers and the humanities*, 31 (1998), pp. 285-300.

Gregory, I., 'Using Geographical Information Systems to Explore Space and Time in the Humanities', in M. Greengrass and L. Hughes (eds.), *The Virtual Representation of the Past* (UK and USA: Ashgate, 2008), pp. 135-146.

Gregory, I.N., and P.S. Ell, *Historical GIS. Technologies, Methodologies and Scholarship* (Cambridge, New York and elsewhere: Cambridge University Press, 2007).

Groza T., *Advances in Semantic Authoring and Publishing* (Amsterdam: AOS Press, 2012).

Grusin, R., 'The Dark Side of the Digital Humanities: Dispatches from Two Recent MLA Conventions', *Journal of Feminist Cultural Studies,* 25:1 (2014), pp. 79-92.

Harpring, P., *Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works* (Los Angeles: Getty Research Institute, 2010).

Harvey, C., and J. Press, *Databases in Historical Research. Theory, Methods and Applications* (New York: Palgrave Macmillan, 1996).

Hellinga, L., 'The Gutenberg Revolutions', in S. Eliot and J. Rose (eds.), *A Companion to the History of the Book* (USA, UK, Australia: Blackwell Publishing, 2007), pp. 207-219.

Hennessey, J. and S.X. Ge, 'A cross disciplinary study of link decay and the effectiveness of mitigation techniques', *BMC Bioinformatics*, 14 (2013), n.pag. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S14-S5> (29 April 2016).

Hillesund, T., 'Digital reading spaces: How expert readers handle books, the Web and electronic text', *First Monday*, 15:4 (2010), n.pag. <http://firstmonday.org/article/view/2762/2504> (23 May 2016).

Hyvönen, E., *Publishing and Using Cultural Heritage Linked Data on the Semantic Web* (California: Morgan & Claypool, 2012).

'Isak Dinesen', Encyclopedia Britannica. Britannica Academic, n.pag. <http://academic.eb.com/EBchecked/topic/163827/Isak-Dinesen> (24 April 2016).

Jabr, F., 'The Reading Brain in the Digital Age', *Scientific American*, 11 April 2013, n.pag. <http://www.scientificamerican.com/article/reading-paper-screens/> (1 April 2016).

Jaffe, K., ‘Social and Natural Sciences Differ in Their Research Strategies, Adapted to Work for Different Knowledge Landscapes’, *PloS one*, 9:11 (2014), n.pag. <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0113901> (13 June 2016).

Jankowski, N.W., et al., ‘Enhancing Scholarly Publications: Developing Hybrid Monographs in the Humanities and Social Sciences’, n.pag. <http://ssrn.com/abstract=1982380> (28 April 2016).

Kirschenbaum, M., ‘What is Digital Humanities and What’s It doing in English Departments?’, in M.K. Gold (ed.), *Debates in the Digital Humanities* (Minneapolis and London: University of Minnesota Press, 2012), p. 3-11.

Larivière, V., et al., ‘The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities’, *Journal of the American Society for Information Science and Technology*, 57:8 (2006), pp. 997-1004.

Lindquist, T., et al., ‘Using Linked Open Data to Enhance Subject Access in Online Primary Sources’, *Cataloging & Classification Quarterly*, 51 (2013), pp. 913-928.

MacMillan, D., ‘Data Sharing and Discovery: What Librarians Need to Know’, *The Journal of Academic Librarianship*, 40:5 (2014), pp. 541-549.

Méndez, E., ‘Metadata Typology and Metadata Uses’, in M. Sicilia (ed.), *Handbook of Metadata, Semantics and Ontologies* (Singapore and Hackenstack, N.J.: World Scientific Publishing Company, 2013), pp. 9-39.

Meroño-Peñuela, A., et al., ‘Semantic Technologies for Historical Research: A Survey’, *Semantic Web*, 6:6 (2014), n.pag. <http://content.iospress.com/articles/semantic-web/sw158> (28 April 2016).

Meroño-Peñuela, A., ‘Semantic Web for the Humanities’, in P. Cimiano et al. (eds.), *The Semantic Web. Semantics and Big Data* (Berlin and Heidelberg: Springer, 2013), pp. 645-649.

Moretti, F., *Distant reading* (London and New York: Verso, 2013).

Mrva-Montoya, A., ‘Beyond the monograph: Publishing Research for Multimedia and Multiplatform Delivery’, *Journal of Scholarly Publishing*, 46:4 (2015), pp. 321-342.

Mueller, M., ‘Digital Shakespeare, or towards a literary informatics’, *Shakespeare*, 4:3 (2008), pp. 284-301.

‘OCLC signs agreements with publishers in the Humanities, Social Sciences and Business’, OCLC, 14 April 2015, n.pag. <https://www.oclc.org/en-CA/news/releases/2015/201512dublin.html> (28 February 2016).

Patterns of information use and exchange: case studies of researchers in the life sciences (UK: The Research Information Network, 2009) <http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf> (8 June 2016).

Pattuelli, M.C., 'Modeling a Domain Ontology for Cultural Heritage Resources: A User-Centered Approach', *Journal of the American Society for Information Science and Technology,* 62:2 (2011), p. 314-342.

Pinker, S., *The Sense of Style. The Thinking Person's Guide to Writing in the 21st Century* (UK and more: Penguin Books, 2015).

Pitti, D.V., 'Designing Sustainable Projects and Publications', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, UK and Australia: Blackwell Publishing, 2004), pp. 471-487.

Ramsay, S., 'Databases', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, UK and Australia: Blackwell Publishing, 2004), pp. 177-197.

Ray, B.C., 'Teaching the Salem Witch Trials', in A.K. Knowles (ed.), *Past Time, Past Place: GIS for History* (California: ESRI Press, 2002), pp. 19-33.

Register, R., *The Essential Guide to Metadata for Books* (New York: F+W Media, 2013).

Reinventing research? Information practices in the humanities (UK: The Research Information Network, 2011) <http://www.rin.ac.uk/system/files/attachments/Humanities_Case_Studies_for_screen_2_0.pdf> (5 March 2016).

Report on Enrichment and Evaluation (Den Haag, Netherlands: Europeana, 2015) <http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf> (2 July 2016).

Rouse, M.A. and R.H. Rouse, *Authentic Witnesses: Approaches to Medieval Texts and Manuscripts* (Notre Dame, Indiana: University of Notre Dame Press, 1991).

Rowe, K., 'Living with digital incunables, or a "good-enough" Shakespeare text', in C. Carson and P. Kirwan (eds.), *Shakespeare and the Digital World. Redefining Scholarship and Practice* (UK: Cambridge University Press, 2014), pp. 144-159.

Scaglione, A.D., *Nature and Love in the Late Middle Ages* (Berkeley and Los Angeles: University of California Press, 1963).

Schöch, C., 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities* 2:3 (2013), n.pag. <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> (2 April 2016).

Seringhaus, M.R., and M.B. Gerstein, 'Publishing perishing? Towards tomorrow's information architecture', *BMC Bioinformatics*, 8:17 (2007), n.pag. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-17> (12 March 2016).

Shotton, D., et al., 'Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article', *PLoS Computational Biology*, 5:4 (2009), n.pag. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663789/> (5 July 2016).

Sperberg-McQueen, C.M., 'Classification and its Structures', in S. Schreibman, R. Siemens and J. Unsworth (eds.), *A Companion to the Digital Humanities* (USA, UK and Australia: Blackwell Publishing, 2004), pp. 161-176.

Tenopir, C., R. Volentine and D.W. King, 'Article and book reading patterns of scholars', *Learned Publishing*, 24:4 (2012), pp. 279-291.

Thaller, M., 'Which? What? When? On the Virtual Representation of Time', in M. Greengrass and L. Hughes (eds.), *The Virtual Representation of the Past* (UK and USA: Ashgate, 2008), pp. 115-124.

Thompson, J.B., *Books in the Digital Age. The Transformation of Academic and Higher Education Publishing in Britain and the United States* (UK and USA: Polity Press, 2005).

Tivnan, T., 'E-book sales abate for Big Five', 29 January 2016, n.pag. <http://www.thebookseller.com/blogs/e-book-sales-abate-big-five-321245> (19 March 2016).

Unsworth, J., 'Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?', 13 May 2000, n.pag. <http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html> (26 May 2016).

Van der Weel, A., *Changing our textual minds. Towards a digital order of knowledge* (Manchester and New York: Manchester University Press, 2011).

W3C, 'RDF', <https://www.w3.org/2001/sw/wiki/RDF> (1 July 2016).

W3C, 'Semantic Web Activity', <https://www.w3.org/2001/sw/> (11 March 2016).

Watson, C.B., *Shakespeare and the Renaissance Concept of Honor* (United States: Princeton University Press, 2015).

'What are Incunabula?', *Incunabula. Dawn of Western Printing* <http://www.ndl.go.jp/incunabula/e/chapter1/index.html> (18 March 2016).

Wiberley, Jr., S.E., 'Subject Access in the Humanities and the Precision of the Humanist's Vocabulary', *The Library Quarterly,* 53:4 (1983), pp. 420-432.

Wierzbicka, A., 'Defining "the humanities"', *Culture & Psychology* 17:1 (2011), pp. 31-46.

Williams, P., et al., 'The role and future of the monograph in arts and humanities research', *Aslib Proceedings*, 61:1 (2009), pp. 67-82.

Woody, W.D., D.B. Daniel and C.A. Baker, 'E-books or textbooks: Students prefer textbooks', *Computers & Education* 55:3 (2010), pp. 945-948.

Woutersen-Windhouwer, S. et al., *Enhanced Publications. Linking Publications and Research Data in Digital Repositories* (Amsterdam: Amsterdam University Press, 2009).

Unpublished secondary sources

Anderson, R., 'Monographs as Essays, Monographs as Databases: Or, the Irrelevance of Authorial Intent', [from an upcoming issue of the journal *Against the grain*], n.pag.

Crossick, G., 'Why Monographs Matter', [from an upcoming issue of the journal *Against the grain*], n.pag.

Web sites

DBpedia, <wiki.dbpedia.org>.

GeoNames, <http://www.geonames.org/>.

Hathi Trust Digital Library, <https://www.hathitrust.org/bib_specifications/>.

Incunabula. Dawn of Western Printing, <http://www.ndl.go.jp/incunabula/e/index.html>.

IUPAC Gold Book, <http://goldbook.iupac.org/>.

NCBI GenBank, <http://www.ncbi.nlm.nih.gov/genbank/>.

Paul Mellon Centre, <http://www.paul-mellon-centre.ac.uk/collections/archive-collections/william-roberts/>.

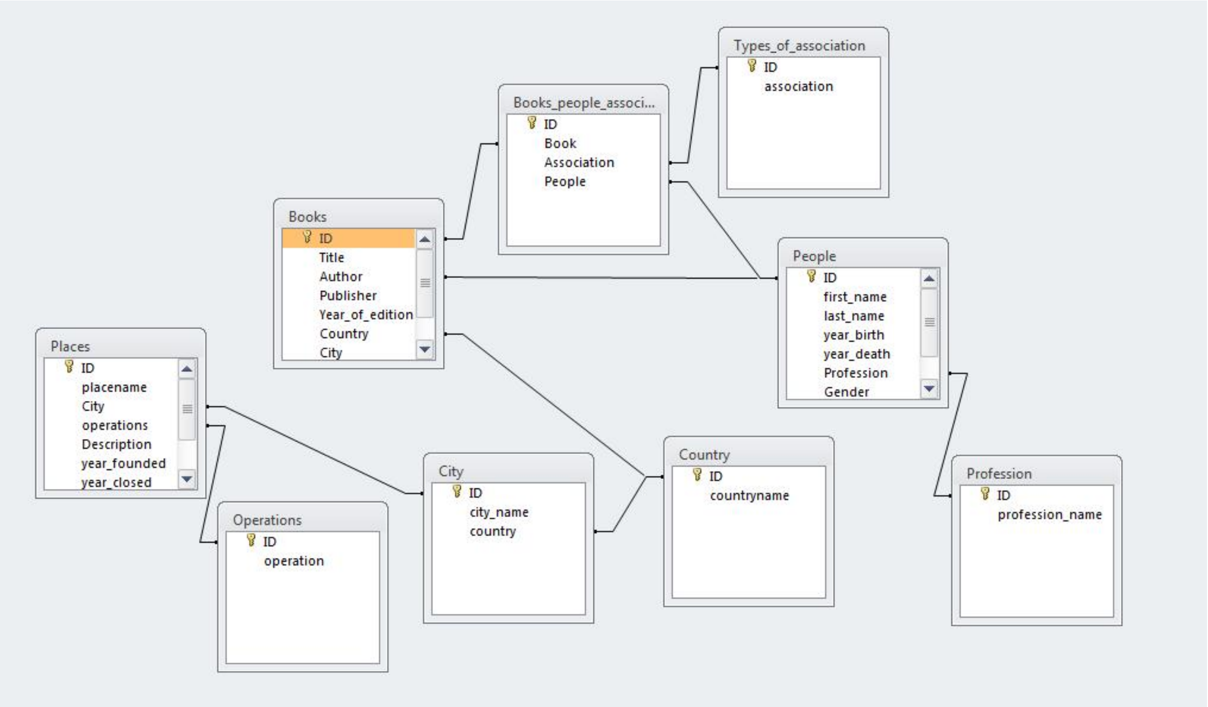
Proteomics Standards Initiative, <http://www.psidev.info/index.php?q%20=%20node/31/>.

PubMed Central, <http://www.ncbi.nlm.nih.gov/pmc/>.

Universiteit Leiden, <http://catalogue.leidenuniv.nl/>.

WorldCat, <https://www.worldcat.org/advancedsearch/>.

Appendix : Creating the web site



The image demonstrates the precise structure of the relational database created for *The Book-hunter in London*. In order to give some idea about the process of creating the database, the base for most of the enhancements discussed in the thesis, I will briefly review the procedure of making it, with a focus on factors which slowed down the process.

The website was created in co-operation with my thesis supervisor, Peter Verhaar, who did all the coding for the website, while my part was the creation of the database and the design of the website's enhancements. The most essential part of creating the enhancements is linking entities to specific passages in the text. If entries in the database are not linked to the text, they are floating in thin air outside of it and lose their value to the user.

Problems with regard to creating these connections limited the possibilities of *The Book-hunter* web site. This was partly due to time constraints, as well as my inexperience with some elements of the enrichment process, but also due to the technical constraints of the tools we had at our disposal. We managed to automatically retrieve a great deal of *names of people* and *names of books*, which were extracted from the index, but the problem was connecting these entities to the main text. Because an entities' name was the only available identifier which could be used to link that entity to the text, any alteration in spelling during the process of editing the database meant the algorithm couldn't make a connection. More importantly, the name of the entity was sometimes spelled differently in the main text than in the index, which also hindered connection with the main text. Therefore, lists of data entries which were retrieved automatically from the index would have needed a sophisticated algorithm capable of linking these entries from the index to identical or similar strings in the main text, and ideally also capable of 'expecting' certain entities on certain pages in accordance with the information in the index.

My experience on this project demonstrated to me that a central part of making semantic enhancement practically viable is minimizing the involvement of a human editor, whose involvement is nevertheless essential. The possibilities of automatic retrieval of information from a text should be maximized, particularly in a commercial setting.[195] The algorithm used for automatic extraction of entities should be looking for matches in a list of *controlled vocabularies*. A match with a controlled vocabulary is important because if a detected entity in the text is *not* derived from an authoritative list, this means both considerably more manual work and diminished possibilities for semantic connections with other publications (see pages 36-37). Without the automated detection and controlled vocabulary, the process becomes very time consuming (as it was in the case of the *Book-hunter* database) and therefore probably unviable commercially. If these two prerequisites are in place, it becomes close to the equivalent of adding another round of editing.

The final essential component which I felt was missing during our coding process was a program that would allow the editor to read through the text and at the same time to quickly and efficiently examine or create the semantic connections *in a single window*. The manual work done for the *Book-hunter* database using *Microsoft Access* meant constant shuffling between the text of the book, the database, and other information sources. This probably slowed the process down by days. If an efficient workflow is in place for these initial steps, creating applications with the resulting data becomes relatively straightforward.

What this experience goes to show is that although there certainly are ways of making the process of enhancing humanities texts efficient, there needs to be a quite extensive framework in place for the results to be completely comprehensive and precise.

[195] The most recent detailed report on workflows for semantic enrichment of cultural heritage is *Europeana*'s: *Report on Enrichment and Evaluation* (Den Haag, Netherlands: Europeana, 2015) <http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf> (2 July 2016).