Affordances and limitations of algorithmic criticism

Verhaar, P.A.F.

Chapter 4

Research data of literary informatics

4.1. Introduction

Literary informatics aims to fuse the scholarly objectives of literary studies with the affordances of digitisation, and concentrates, additionally, on the epistemological consequences of adopting computational approaches. In a minimalist scenario, the new methods are used principally to serve existing objectives and to rationalise established heuristics. David Berry stresses, nevertheless, that digital methods "have profound effects on all aspects of the disciplines".[268] Computation is undeniably characterised by a particular logic, and the use of technology demands a full or a partly subjection to this logic. Computer-based scholarship is initially a hybrid form of scholarship, in which parts of the conventional aims and methods are merged with the modus operandi that follows from technical exigencies. The precise outcome of the encounter between the digital and the traditional are often difficult to predict. It seems clear, nevertheless, that a careful examination both of the nature of digital methods and of the needs of the scholarly field in which these methods are adopted can help to identify potential incompatibilities, as well as opportunities for a productive confluence.

The two critical methods which are contrasted in this thesis have been discussed, to some extent, in the previous two chapters. Chapter 2 of this thesis has described conventional approaches to studying poetry, and Chapters 3 has explained some of the ways in which literary texts can be analysed via digital means. Chapter 3 has characterised the field of literary informatics in practical terms, concentrating on concrete tools and on specific research projects. This chapter seeks to characterise the nature of literary informatics research on a more conceptual level. One crucial characteristic of computer-based scholarship in general is that it is based on data. Digital humanities research often begins with a process in which the artefacts that are studied are converted into discrete data values, and the eventual scholarly claims are commonly based on statistical analyses of these data sets. Despite the general importance of the concept of research data, which Christine Borgman has referred to as "the foundation of scholarship",[269] it can be observed that there is still a degree of uncertainty about

[268] David M. Berry, "Introduction", in: *Understanding Digital Humanities*, New York: Palgrave Macmillan 2012, p. 13.

[269] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, p. 115.

the precise meaning of the term within humanistic research.[270] It may have different meanings within the various branches of the humanities, and it may confusingly refer to different types of scholarly artefacts. This chapter introduces a number of terms that can be used to describe generic aspects of research data, and provides a definition of the various data types that are used and created by scholars. This framework is subsequently used to characterise the data that are produced by scholars in the field of literary informatics.

4.2. Definitions of research data

In recent years, there has been a growing interest in the potential advantages of the curation of research data among researchers, politicians and funding agencies, and whereas data, as noted, is a broad and a convoluted term, a number of generic definitions have been proposed. Many of these agree that data can have evidentiary value and that they can be used to support or to validate particular claims. The OECD report on *Principles and Guidelines for Access to Research Data from Public Funding* defines data as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings".[271] The data policy of the University of Edinburgh stresses similarly that data, "unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results".[272] Research data are often considered to be a key element in the "chain of evidence" that underlies scholarly research in all disciplines.[273] As data commonly serve as the building blocks which ultimately enable scholars to construct an argumentation, they are generally viewed as semi-manufactures, rather than as the final products of a research project.

The definitions that have been cited can help to clarify the rationale of research data, but they do not offer a precise description of their nature. Definitions which focus more closely on their essence often emphasise that data are primarily descriptions or representations of the objects in a particular domain. Dervos and Coleman, for instance, discuss a distinction between 'facts' and 'data'. They explain that the term "fact" refers to "things done, that is, deeds or acts made into

[270] Deploring the confusion which frequently surround the term "research data" within the context of the humanities, Borgman urged digital humanists to provide better descriptions of the ways in which data are created, evaluated and used. See Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities", s. 23.

[271] *OECD Principles and Guidelines for Access to Research Data from Public Funding*.

[272] "Research Data Management Guidance", <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt> (22 June 2014).

[273] Marlo Welshons, *Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*, (2006), p. 18.

something known" and to phenomena or events which "pertain to objective reality".[274] Data, on the other hand, "represent real world facts". They comprise "the outcome of measurements conducted in relation with real world phenomena".[275] A kindred view is offered by Marcia Bates. In an attempt to define data from a biological and evolutional perspective, Bates describes data as "that portion of the entire information environment available to a sensing organism that is taken in, or processed, by that organism".[276]

In the academic field of information science, many theorists view data as one layer in a larger hierarchy which connects the term to the related concepts of "information", "knowledge" and "wisdom". The series of models which represent the associations between these four concepts have been referred to varyingly as the "DIKW Pyramid" or the "DIKW Hierarchy". The model has undergone a number of revisions and extensions since its initial conception in the late 1980s,[277] but it is generally maintained that each layer derives from an underlying layer, and that the base of the hierarchy is ultimately formed by data. In most interpretations of the model, data are conceptualised as factual and non-interpreted representations of specific external phenomena. Russell Ackoff describes data as "symbols that represent properties of objects, events and their environments",[278] and Lieuw views data as "the storage of intrinsic meaning, a mere representation" whose main purpose is "to record activities or situations, to attempt to capture the true picture or real event".[279] In agreement with the etymology of the word "data", the theories and definitions that have been cited claim that data can be acquired by observing objects or events within an environment that is largely "given" a priori.

Definitions which depict data as representations of existing phenomena can help to embed the term within an existing discourse about the nature of the digital humanities. Many scholars view digital humanities as a field that is centrally concerned with the development of and the critical engagement with models. In his monograph *Humanities Computing*, Willard McCarty argues that the computer's value to humanities research stems primarily from the capacity to support "the

---

[274] This definition is quoted by the authors from Lawrence McCrank, *Historical Information Science* (Medford: Information Today, 2002), p. 627.

[275] Dimitris A. Dervos & Anita Sundaram Coleman, "A Common Sense Approach to Defining Data, Information, and Metadata", in Gerhard Budin, Christian Swertz, & Konstantin Mitgutsch (eds.)*Proceedings of the Ninth International ISKO Conference*, (Würzburg: Ergon-Verlag, 2006), pp. 51–58, p. 55.

[276] Marcia J. Bates, "Information and Knowledge: An Evolutionary Framework for Information Science", in: *Information Research*, 10:4 (2005), n.pag.

[277] For a good discussion, see J. Rowley, "The Wisdom Hierarchy: Representations of the DIKW Hierarchy", in: *Journal of Information Science*, 33:2 (February 2007).

[278] R. L. Ackoff, "From Data to Wisdom", in: *Journal of Applies Systems Analysis*, 16 (1989), p. 3.

[279] Anthony Liew, "Understanding Data, Information, Knowledge And Their Inter-Relationshipso Title", in: *Journal of Knowledge Management Practice*, 8:2 (2007), n.pag.

heuristic process of constructing and manipulating models".[280] McCarty defines a model as "a representation of something for purposes of study, or a design for realizing something new".[281] John Unsworth stresses likewise that computer-based humanistic research is fundamentally "a practice of representation" or "a form of modelling". In this form of research, "the computer is used as a tool for modelling humanities data and our understanding of it".[282] Humanities scholars are mostly interested in the nature, the historical development or the reception of cultural or artefacts, but these original artefacts are often in a format which complicates or even precludes a systematic digital analysis. When scholars aim to investigate texts which originated as physical objects, the relevant properties of these objects obviously need to be represented via bits before these can be studied computationally. Berry explains that the digital humanities "try to take account of the plasticity of digital forms and the way in which they point toward a new way of working with representation and mediation, what might be called the digital 'folding' of reality".[283] The idea that the concept of representation is crucial to the definition of data is also underscored in the Digital Humanities Manifesto 2.0, which was authored by Pressner and Schnapp. The text characterises the digital humanities as an area of research which "values the copy more highly than the original". Interestingly, the authors exploit the etymology of the word "copy", whose original meaning of "abundance" survives in the word "copiousness", to stress the extensive and widespread availability of digital surrogates of cultural artefacts.[284]

Computer-based literary research engages with text in a manner that is distinctly circuitous. Works of literature clearly form the ultimate objects of research, but critical analyses focus primarily on digital surrogates of these works. The DIKW model sets forth the view that data consist of surrogates of objects or of events, and this expedites the application of the term within the context of humanities research. Other characteristics of the DIKW model also complicate its pertinence to humanistic research, however. In a critique of the DIKW theory, Martin Frické writes that the model has a bias towards the natural sciences, and that it operates exclusively within the confines of positivism or empiricism. The model is based on the incorrect assumption that knowledge can only be obtained through a systematic analysis of sensory or observational data. The DIKW hierarchy allows no room for "unobservable ('theoretical') entities and properties"

[280] Willard McCarty, *Humanities Computing* (Basingstoke; New York: Palgrave Macmillan 2005), p. 23.
[281] Ibid., p. 24.
[282] John Unsworth, "What Is Humanities Computing and What Is Not?", in: Melissa Terras, Julianne Nyhan, & Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, 2013, pp. 36–37.
[283] David M. Berry, "Introduction", p. 2.
[284] Jeffrey Schnapp, Peter Lunenfeld & Todd Pressner, *The Digital Humanities Manifesto 2.0*, p. 14. The original text uses upper case for some of the central nouns, but the capitalisation has been removed in this quotation.

nor for objects or phenomena in the physical domain "for which no instruments of measurement exist".[285] Humanities scholars do not necessarily base their argumentation on factual or objective properties of observable phenomena, and the observations that are made by scholars are frequently of an interpretative or of a speculative nature. For this reason, conceptualisations of humanistic data can only be cogent when they acknowledge that data can be either factual or interpretative representations of physical or born-digital cultural artefacts.

Kitchin discusses a useful difference between "captured data" and "derived data". The first term refers to raw and unprocessed data which are generated "through some form of measurement such as observation, surveys, lab and field experiments, record keeping [...], cameras, scanners and sensors".[286] Captured data record neutral or unprocessed facts about an observable reality. Unprocessed data-sets are often in a format that cannot be queried systematically, and, if this is the case, researchers need to restructure, classify or normalise the data.[287] Derived data, in contrast, are "produced through additional processing or analysis of captured data". In general, the level of neutrality or objectivity decreases with each phase of further processing.[288]

Building on a conceptual description provided by Davis et al., Unsworth identifies a number of generic characteristics of data. For the current discussion, three properties are of particular relevance. Data, firstly, are created to enable or to expedite particular types of analyses. They imply a "fundamental conception of intelligent inference",[289] and they sanction or recommend specific types of computational manipulations. Secondly, as is also stressed by McCarty, digital representations are typically based on an ontology. In philosophy, the term

[285] Martin Frické, "The Knowledge Pyramid: A Critique of the DIKW Hierarchy", in: *Journal of Information Science*, 35:2 (21 November 2008), p. 4.

[286] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences* (2014), p. 7.

[287] Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, (Bath: 2007), p. 15.

[288] Research data can be categorised in many other ways. Kitichin discusses a distinction between primary, secondary and tertiary data. Primary data are "generated by a researcher and their instruments within a research design of their making". Secondary data, on the contrary, are "data made available to others to reuse and analyse that are generated by someone else". Tertiary data, finally, result from calculations and other forms of processing. Examples include "counts, categories, and statistical results". A similar classification is provided by Norman Blaikie, who argues that the distinction between primary, secondary and tertiary data also arranges the data according to the "notion of distance from the phenomena". Such distance is minimal with primary data, as they are "the result of direct contact between the researcher and the source". Tjalsma and Rombouts use the terms "primary data" and "secondary data" to distinguish between "empirical, observed, measured data" on the one hand, and "data derived from sources created previously". This description differs, nevertheless, from the definitions which were given by Kitchin and by Blaikie. Because of these differences he distinction between primary, secondary and tertiary data will not be used in this thesis.

[289] John Unsworth, "What Is Humanities Computing and What Is Not?", pp. 40–41.

"ontology" is generally used to refer to the "science or study of being",[290] but, in the context of humanistic modelling, the term is used more restrictively to describe the aspects of the original which are reproduced in the digital surrogate. It is an abstract theoretical conceptualisation of the original object to be modelled, enumerating the various qualities that need to be represented. A representation can never be fully representative, and it is inevitably based on a prior identification of the characteristics that are considered essential. As such decisions simultaneously imply a statement that the remaining aspects may be ignored, models "inevitably lie, by omission at least".[291] A third characteristic of digital surrogates is that they make use of a formal representation language. This language may also be referred to as a data format. Formatting is a "mechanism for describing data, i.e., for mapping concepts of a data model to digital objects such as files or memory".[292] A data format ensures that the aspects and the concepts which are considered relevant can be captured and manipulated on a digital device. The formal representation language consists of a range of symbols that can be used to express particular concepts, and a syntax that prescribes the manner in which these symbols may be combined into valid statements or data structures. Examples of data formats include plain texts, TEI-encoded texts, images, RDF-based annotations and formats in relational databases.[293]

This chapter provides a classification of the different types of research data that can be created and reused within literary informatics research. This classification is based on a consideration of the ontologies that underlie the various data formats. Ontologies and data formats are strongly linked, nevertheless. Because a particular data format can never represent all textual aspects, the choice of a format also implies specific ontological commitments. The nature of the ontology that underpins a particular surrogate can be investigated by comparing the properties of the surrogate to the qualities of the original. As a preamble to the characterisation of these data formats, the following section provides a detailed and generic discussion of the various aspects of the literary works that can be studied. The generic and expansive ontology that is presented in the next section can be used to analyse the more limited ontologies that are maintained within individual data formats.

Following a definition provided by Borgman, the original works that are studied can be referred to as sources. Sources, more specifically, comprise all the relevant materials that were created outside of the context of the research process, and which form the input for scholarly enquiry. Sources need to be distinguished from resources, which are the "data, documents, collections, or services that meet

[290] "Ontology", in *Oxford English Dictionary*, <www.oed.com> (16 May 2015)
[291] R. Davis, H. Shrobe & P. Szolovits, "What Is a Knowledge Representation?", in: *AI Magazine*, 14:1 (1993).
[292] Robert E. McGrath, *XML and Scientific File Formats*, (Urbana-Champaign: 2003), p. 6.
[293] These formats are discussed in more detail in the following sections.

some data or information need".[294] Resources are created by scholars themselves in the course of their research. The nature of research data can understood by considering their relation to the sources they mimic.

4.3. Sources

Some of the generic properties of textual sources can be analysed effectively by making use of the conceptual entity relationship model which was developed by the *International Federation of Library Associations* (IFLA), under the name *Functional Requirements for Bibliographical Records* (FRBR).[295] The FRBR model, which was devised originally to represent the various bibliographic levels which may be present in a library catalogue,[296] distinguishes four cardinal terms. A "work", firstly, is "a distinct intellectual or artistic creation". It can be understood as a Platonic representation of a particular creation, because "there is no single material object one can point to as the work". The second entity in the FRBR model is the "expression". It is an "intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object movement, etc., or any combination of such forms". The original text as written by the author is considered to be an expression. The sequence of characters that is produced is an embodiment of a particular work. Subsequent new editions or translations of this text establish a new sequence of characters, and they are consequently new expressions of this work. A 'manifestation', thirdly, is an expression which has been presented on a particular medium, using a specific typography.[297] When a single edited version is made available multiple times with a different typographical appearance or on different media, these are all considered to be different manifestations. An 'item', finally, is "a single exemplar of a manifestation". The 'item' is the only concrete class of objects in the FRBR model, since the three additional levels are abstract concepts which can be used to make statements about the ways in which items can be connected. Digital surrogates of printed works are necessarily representations of a particular item. Such an item contains a particular string of characters, and it is presented through a particular typography.

Whereas the FRBR model can be applied effectively to clarify the differences between separate printed editions, the conceptualisation may be less appropriate for literary texts that are transmitted via other media. The distinction between expression, manuscripts and item is problematic, for instance, for literary texts

[294] Christine Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, p. 122.
[295] *Functional Requirements for Bibliographic Records: Final Report*, (2009).
[296] Karen Coyle, "FRBR, the Domain Model", in: *Library Technology Reports*, 2010, p. 21.
[297] Views vary on whether or not a change in medium actually causes the creation of a new manifestation.

which were created before the invention of moveable type, and which have survived as handwritten sources. When scribes copied a manuscript, they inevitably introduced alterations in the text, thus establishing a new expression. As each copy was produced in the hand of an individual scribe,[298] each new manuscript may also be seen as a separate manifestation. Thirdly, each manuscript is also unique, and this renders the concept of the item, as a level distinct from the manifestation, inconsequential. In the case of a manuscript, only two levels can reasonably be distinguished, since the concepts of expression, manifestation and item coalesce. The distinction between manifestations and items may also be disputed in the case of electronic sources. Van der Weel notes that the essential virtuality of digital files admits the possibility to produce an unlimited number of copies which are indistinguishable from the original. The copy and the original are so much alike, in fact, that it may no longer be appropriate to refer to the duplicated file as a copy.[299] The concept of the copy appears to be salvaged, however, by recent technical developments. A growing number of applications enable readers to annotate and to manipulate specific copies of digital resources.[300] The personal annotations which were added by a particular reader, and which are clearly unique to a particular file, can be relevant from a scholarly perspective. In the case of such annotated digital files, reinstating the item as a separate entity can be justified.

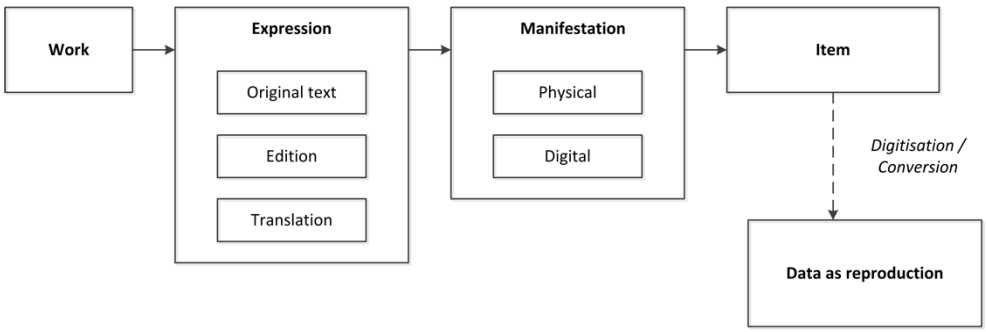

Image 4.1. Entities of the FRBR model

[298] In scriptoria which had implemented the pecia system, a single text can also be written in the hands of multiple scribes.

[299] Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge*, p. 150.

[300] Applications such as iAnnotate, Browzine and Mendeley form important examples.

Whereas the different levels that are proposed by FRBR cannot be applied equally suitably to all the sources that can be investigated by literary scholars, the model's main value for the current discussion lies in the fact that it identifies a number of crucial analytic components of the concrete physical or digital object whose properties are mimicked in a surrogate. A first important observation is that a textual source, as an expression of a work, comprises a particular sequence of characters. Peter Shillingsburg refers to these characters as the "lexical codes", and explains that these consist of the "the more usually acknowledged aspects of a text, letters, accents, and punctuation".[301]

FRBR also draws attention to the fact that the original artistic expression can be extant in different editions. Literary research can incidentally take place by inspecting the original manuscripts of an author, but the majority of critics base their findings on published editions of literary works. While, in the natural sciences, research typically concentrates on phenomena which are given and which exist independently of the observer, literary scholars can intervene directly in the domain they study by actively shaping the nature of textual sources. The aim of editing, broadly speaking, is to make texts more accessible or more useful to specific audiences by organising, correcting or even paraphrasing original works. Scholarly editing is a specific type of editing in which the various modifications that are made to the primary sources are based on scholarly research. The type of research that underpins such scholarly editions is commonly referred to as textual criticism.[302] The alterations that are introduced mostly follow the rules and the standards that are prevalent within specific editorial frameworks or traditions.[303] Elena Pierazzo explains that the different approaches that are followed in scholarly editing are defined, amongst other aspects, by "the way they handle the evidence offered by primary sources", by "the way they reconcile contrasting readings from different sources" and by "the importance given to authorial intention".[304]

Tim Machan discusses a useful distinction between lower criticism and higher criticism. The former type of criticism mostly entails "the establishment of literary, social, and cultural contexts and thus subsumes biography, bibliography, and

[301] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 16.
[302] Kathryn Sutherland, "Being Critical: Paper-Based Editing and the Digital Environment", in: Marilyn Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate, p. 13.
[303] Examples of such frameworks include stemmatics, which is frequently the norm for editions of medieval manuscripts, Walter Greg's copy-text theory, which has often been applied to early modern print materials, and genetic criticism, which aims to represent the writing process that was followed by an author. See Elena Pierazzo, *Digital Scholarly Editing: Theories, Models and Methods* (Farnham: Ashgate 2015), p. 11.
[304] Ibid.

textual criticism". [305] Higher criticism, by contrast, is directed more specifically towards explanation and literary interpretation. Lower Criticism is "commonly viewed as the more factual or 'scientific'" approach as it establishes "numerical, analytical, and categorical information which is used to define historical realities".[306] Machan adds that there is typically a reciprocal relation between textual criticism and interpretative research. The results of the exertions of editors and textual critics form the sources of the literary critic. Textual criticism aims to produce the semi-manufactures on which literary critics can base their arguments, and the lower form of criticism serves as "a stepping stone to the Higher, and sometimes more important, one".[307] Conversely, the needs and the aims of literary critics also provide a context for the activities in the field of lower criticism. Particular interpretations of a text may affect the way in which textual critics approach later editions of this text.

Historical-critical editions mostly aim to represent the author's original intentions as directly as possible, and to avoid any bias on the part of the editor. Many of the core tenets of the historical-critical approach were originally formulated by the philologist Karl Lachman, who proposed a stemmatic approach aimed at tracing the lineage between the various witnesses of a text. The objective of the approach, ultimately, is to expose the work in its original form, free from errors and omissions.[308] In a similar vein, Walter Greg proposed that critical bibliography should serve as an anchor for the subjective activities of the literary critic, and stressed that editors should refrain as much as possible from making critical interpretations.[309] More recent theorists of the field of textual criticism have stated that such neutrality is unfeasible, and have embraced the fact that editions invariably reflect the preferences of textual critics. The objective to reconstruct an original and authoritative version of a text has now largely been replaced by the conviction that all extant versions of the text have value in themselves. The New Bibliography movement, for instance, abandoned the notion of textual idealism, and recognised that editors generally write much of their own interpretations into a text. Paul Zumthor's influential "mouvance" theory suggested likewise that all edited texts should be placed in their social context, as each textual variant is part of the reception history of the text. Machan emphasises that "there can be no value-free textual criticism", and Peter Shillingsburg agrees that "the compilation of a scholarly edition is the interpretive best thinking of an editor and is not the

[305] Tim William Machan, "Late Middle English Texts and the Higher and Lower Criticisms", in: Tim William Machan (ed.), *Medieval Literature: Texts and Interpretation*, Binghamton: Center for Medieval and Early Renaissance Studies 1991, p. 4.

[306] Ibid.

[307] Ibid., pp. 4–6.

[308] Paul Maas, *Textual Criticism* (Oxford: Clarendon Press 1958), p. 3.

[309] G Tanselle, *Textual Criticism since Greg : A Chronicle, 1950-1985* (Charlottesville: University Press of Virginia 1987).

establishment of a text for all time".[310] Since literary works are typically available in multiple editions, and since each edition, to a higher or a lesser degree, entails a particular reading of a text, literary scholars who aim to study a work digitally need to be critical with respect to the edition they aim to model.

On paper, an edition generally publishes a single variant of the text, since a paper-based compilation of all variants would be impracticable and economically unviable. In the digital realm, the mouvance theory can manifests itself in practical terms in what Vanhoutte refers to as the "maximal edition".[311] The term refers to an edition in which the entire transmission history of a text is represented, and in which all available variants are included. In most cases, maximal editions can only be created in a digital form. To ensure that users can purposefully navigate a large numbers of variants, maximal editions must exploit the malleability of digital texts, in an interface in which fragments from selected witnesses can be shown side-by-side. According to Sutherland, however, the possibility to include each extant witness only has limited worth. If the editor omits the act of selection, the burden of having to select a specific version shifts to the end-user, who is normally not in a position to evaluate the significance of the various options. For this reason, Sutherland refers to such digital editions as "recyclable wastebanks".[312]

Within a particular expression of a work, a number of textual aspects can be identified. The lexical codes of the text, which are often particular to a given edition, contains a logical structure. The characters that make up a text are divided over distinct logical units such as chapters, sections, paragraphs and sentences. The various logical units are normally used in the service of a rhetorical structure. Units such as paragraphs and section often have a specific function in the narrative or in the overall the argumentation of the text. The rhetorical structure enables readers to trace, for instance, how certain conclusions follow from premises which have been introduced earlier. This logical structure has usually been conceived before the text was cast onto a particular medium. There is also a second set of structural units, however, which is brought into existence by the placement of the full text onto paper. Examples in the latter class include running titles, page numbers and title pages. These components were generally not conceived of by the author, and they have usually been added to facilitate the navigation through the text. Gerard Genette explains that these structural components from part of the paratext. Paratextual units which are visible within the publication, and which

[310] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 171.
[311] Edward Vanhoutte, "Every Reader His Own Bibliographer - an Absurdity?", in: Marylin Deegan & Kathryn Sutherland (eds.), *Text Editing, Print and the Digital World*, Farnham: Ashgate 2009, p. 111.
[312] Kathryn Sutherland, "Being Critical: Paper-Based Editing and the Digital Environment", p. 26.

result from the text's *mise-en-page*, are referred to, more specifically, as the peritext.[313]

The actual running text, not including the peritext, displays a linearity. In this context, the term linearity refers to the notion that the units of the language need to be processed in a fixed order, and that the units, to a large extent, derive much of their meaning from their placement within from this particular context. In his *Course on General Linguistics*, De Saussure has noted that such a linear progression is characteristic of all texts in natural language. He explains that the linguistic signal has a "temporal aspect, and hence certain temporal characteristics: (a) it occupies a certain temporal space, and (b) this space is measured in just one dimension: it is a line".[314] Linearity is characteristic of both written and spoken texts. In the case of a paper-based text, "a spatial line of graphic signs is substituted for a succession of sounds in time".[315] For readers, it is generally difficult to locate units of information instantaneously and separately from their context. To fully understand the meaning of a narrative, readings generally need to consider the broader context.

Next to the fact that a text consists of specific string of lexical codes, the FRBR model also recognises that textual sources have a typography. The function of typography is generally to clarify the logical structure, which includes the components that belong to the peritext. Items such as paragraphs, block quotes, section heading and footnotes are normally rendered distinct visually, and this enables the reader to decipher the logical category of each segment. The nature of the units that belong to the peritext is often clarified via the positioning of these units on the page. Running titles, for instance, can be recognised as such because they are placed at the top of the page.

In FRBR, typography is not considered to be part of the original artistic expression. The definition of the term expression excludes "aspects of physical form, such as typeface and page layout, that are not integral to the intellectual or artistic realization of the work as such". According to the model, the work of the typographer may be compared to that of a conductor of a piece of music, who essentially provides an interpretation of a work that was conceived originally by a composer. If the same content is published multiple times, but with a distinct typography, these two texts still belong to the same expression. A related argument can be found in the article "What is a Text, Really?", in which DeRose et al. consider "the question of essentials: What is it which, if changed, makes a document essentially different, and what is it which can change, yet a document remains 'the same?'". The authors conclude that the words of the text constitute the

[313] Gerard Genette & Marie Maclean, "Introduction to the Paratext", in: *New Literary History*, 22:2 (2010), pp. 263–264.
[314] Ferdinand de Saussure, *Course in General Linguistics* (London: Duckworth 1983), p. 69.
[315] Ibid., pp. 69–70.

genuinely "meaningful units". When the words change, a new text originates, while "adjustments of typography" appear to be "superficial and transient rather than essential".[316]

Various authors have argued, nevertheless, that typography makes a crucial contribution to the overall experience and significance of text. Don McKenzie, for instance, explained that "the material form of books, the non-verbal elements of the typographic notations within them, the very disposition of space itself, have an expressive force in conveying meaning". The strong connection between form and content is underscored by the etymology of the word 'text', which indicates "a process of material construction".[317] In a printed text, content and typography are woven together into a single indelible unit, and the form actively contributes to the production of the text's message. According to McKenzie, texts ought to be studied as "recorded forms" which have originated at a particular location and at a particular time. In printed books, specific aspects of the text's layout are indicative of a particular reading and, in turn, these influence the manner in which new readers interpret the text.

The position that typography is transparent and semantically void is confounded severely by literary texts in which the author has explicitly used the textual form as an expressive element. The typography of a text strongly affects the manner in which readers experience this text. Lennard stresses that decisions about the layout of a text are not made exclusively by printers or by publishers, since authors increasingly begin to employ this aspect in order to convey meaning.[318] Works in the literary genre that is commonly referred to as "concrete poetry", for instance, have been "composed with specific attention to graphic features such as typography, layout, shape, or distribution on the page".[319] Eugen Gomringer's "Silencio" and Decio Pignatari's "Bebe coca cola"[320] exhibit the notion of isomorphism, which is a genre of poetry in which shape and meaning are considered identical. Other examples of texts in which the typography has a clear semantic function can be found in the works of George Herbert, Paul van Ostaijen en Dom Silvester Houedart. In Herbert's pattern poems, for instance, of which "The Altar" and "Easter Wings" are probably most widely known, the shape that is formed by the words on the page ingeniously depict and support the subject matter. In digital surrogates of such concrete or visual poems, it is essential to

[316] Steven J. DeRose et al., "What Is Text, Really?", in: *Journal of Computing in Higher Education*, 1:2 (1990).
[317] Donald Francis McKenzie, *Bibliography and the Sociology of Texts* (Cambridge: Cambridge University Press 1999), p. 3.
[318] John Lennard, *The Poetry Handbook: A Guide to Reading Poetry for Pleasure and Practical Criticism*, p. 47.
[319] *The Princeton Encyclopedia of Poetry and Poetics*, p. 294.
[320] Both poems are discussed in *The Princeton Encyclopedia of Poetry and Poetics*.

ensure that the characteristic aspects of the form can be retained, since these texts would otherwise be bereaved of much of their expressive value.

McKenzie's "recorded forms" consist of lexical codes combined with a particular typography, and, accordingly, they correspond to 'manifestations' in the FRBR model. In the context of the printed book, the expression and the manifestation inevitably coincide in a physical item, as it is not physically possible to separate lexical codes from the typography. Terms which, in the context of physical publications, largely refer to abstract concepts may actually be used to describe distinct types of files in the context of the digital medium. A plain text which consists exclusively of Unicode characters can be considered an expression of a text. Such an expression can be marked up typographically, using, for instance, a DTP program, and the result will then be a manifestation of this text. Similarly, when a text is marked up using XML, a stylesheet can be applied to this document in order to render the encoded content in a specific typographical form. Because of this separation of content and form, readers of digital sources can be confronted with texts in which the typography is intrinsically unstable. A distinction can be made between fixed formats, such as PDF or TIFF, and rescalable formats, such as HTML or EPUB. When texts are disseminated in rescalable formats, layout artists only have limited possibilities to fix the presentation. In the case of HTML, for instance, the typographical appearance often depends for a large part on the particular settings of the browser in which users read a text. In eBooks based on EPUB, users are often able to personalise the presentation of the text, and to change font faces, letter sizes or background colours. As readers can flexibly weave and re-weave the contents into new forms, rescalable eBooks complicate MacKenzie's notion of texts as recorded forms. The plasticity of digital sources undermines the typography's ability to convey and to support a particular reading.

The current enumeration of the cardinal properties of textual sources concludes with a brief mention of a number of obvious characteristics, which need to be mentioned for the sake of completeness. A text, evidently, has semantic contents. The text invariably addresses specific topics, and it frequently references items in specific generic categories of information, such as personal names, geographic terms or book titles. As discussed in the previous chapter, literary texts can also make use of literary devices, such as metre, rhyme, alliteration and imagery. Texts, furthermore, have linguistic properties. Literary writing often employs a register of language that differs in a number of important ways from common colloquial language. Literary analyses may, for this reason, also focus on occurrences of syntactical categories or of particular grammatical constructions. Words, when read aloud, are also associated with sounds. Literary devices are often based on a skilful use of such sounds. Literary research, and, in particular, analyses of poetry, often concentrate on the phonetic aspects of a text. Words in the English language are generally used in conjugations and in declensions. In some cases, it can be useful to replace inflected forms with lemmas, which are their base dictionary forms.

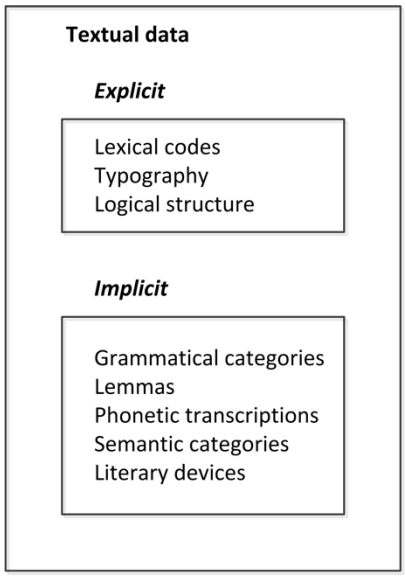

Figure 4.2. Core aspects of texts

The properties that have been discussed in this section, and which may potentially be mimicked in a surrogate, are summarised in figure 4.2. Literary research focuses on texts written by literary authors. Importantly, the verbal construction that is studied is rarely a fully neutral representation of the words that were written by the author, as the nature of the lexical codes is often informed by the editor's opinion about the nature and the purpose of the literary text. Additionally, while the text primarily produces its meaning via the lexical codes, the form of the text frequently contributes strongly to the overall effect of the text. This is particularly the case for concrete or visual poetry. Texts often bear the marks of subjective interpretations performed by textual critics or by typographers. The published edition, nevertheless, is an objective artefact which can serve as the basis for subsequent digital scholarship.

4.4. Captured data

Computer-based literary research typically commences with a process in which a collection of sources which were intended to be read by human beings are transformed into artefacts that can be studied computationally. The fact that digital scholarship demands a conversion is obvious in the case of texts which were initially cast onto an analogue medium. Converting physical sources into digital objects inexorably discards aspects that are tied uniquely to the physical form, such

as the paper quality, the dimensions, the tactile sensation and the smell of the object. Such characteristics can be studied exclusively by consulting the original. A conversion from analogue into digital is obviously not needed for born-digital texts. In most cases, nevertheless, sources which are natively digital still need to be converted, as such sources often contain navigational or typographical aspects which may hinder a systematic analysis. The form that was devised for human readers is not necessarily suitable for scholarly purposes.

This chapter makes a broad distinction between data formats that may be used most appropriately for the representation of observable and explicit aspects of texts on the one hand, and formats that are more suitable for the description of aspects which are implicit or latent in the original works on the other. Data which describe observable aspects of the original object can be classified as "captured data". In the previous section, eight cardinal aspects of texts have been identified, and two of these, the lexical codes and the typography, have an explicit and observable presence. About the text's typography, it must be stressed that, whereas the formal features are manifest, the semantics associated with these presentational devices is implicit. The aim of the typography is mostly to clarify the text's logical structure, but there are no explicit markers which unambiguously identify the nature of the various logical components. A group of characters separated by hard returns may constitute a paragraph, a stanza, an epigraph or a block quote, among other options. In most cases, a human reader can decode the typographical conventions flawlessly, in the same way as he does the lexical codes.[321] Since the nature and the function of these units are not declared explicitly, these cannot be identified directly by computers. Because of the implicit semantics of typography, it may be argued that the logical structure is partly manifest and partly latent.

The plain text format can be used to reproduce the lexical codes of a text. Within this format, the characters that are contained within the work can be accessed separately. In a sense, the plain text is a reversal of the process that was initiated by the invention of moveable type. While Gutenberg used a limited set of cast characters to produce fixed and stable objects, many of the affordances of the electronic text are based on the fact that its letters, digits and punctuation marks can be manipulated separately. As was discussed, machine-readable texts are made available by an increasing number of libraries or commercial organisations. The vast text base that has been assembled as part of *Project Gutenberg*, for instance, is a common source of primary data for textual scholars. About this particular resource, Peter Shillingsburg notes that it often has insufficient information to ascertain that the texts are sufficiently accurate. Additionally, it is frequently

[321] Adriaan Van der Weel explains that human readers "truly deserve to be called *homo typographicus*" because of the "astonishing ease with which we are capable of assessing unconsciously the purport of textual messages without even reading a word of the actual text". See Adriaan van der Weel, *Changing Our Textual Minds : Towards a Digital Order of Knowledge*, p. 69.

unclear whether or not editors have chosen a source text that has a degree of authority or historical importance.[322] If no accurate and authoritative plain texts are available, scholars may produce plain texts themselves, by transcribing the text, or through the use of OCR software.

Captured data are rarely fully accurate and seldom entirely objective. While digitisation projects often strive to produce reliable and unadulterated representations of the original sources, it is usually impossible to avoid the introduction of alterations during the conversion process. The structural inadequacy of the results of OCR scanning, for instance, have been documented extensively.[323] The results can be particularly poor for books containing uncommon font types or for books with a low print quality. In the case of handwritten materials, the texts usually need to be transcribed manually. During such processes, subjectivity can never be avoided entirely. Human scholars may make typing errors unknowingly, but they may also change the text more consciously for the purpose of specific editorial interventions. When a particular hand is difficult to read, making a transcription obviously demands interpretation. In addition, medieval manuscripts often contain various abbreviated words and phrases which transcribers may choose to expand, but views may vary on what these abbreviations actually stand for. In general, a wide range of choices need to be made during the conversion of analogue sources into digital resources, and different persons may also take different decisions.

When texts that have originated on a paper medium are scanned and converted into machine-readable text, this is not a simple migration of content, as the digitised version differs from the original in a number of important ways. Research which is based solely on data in the plain text format crucially disregards the text's typography. As plain texts are essentially immaterial and formless resources, the use of this format is problematic for literary works in which the typographical presentation is an inherent part of the artistic expression. Plain machine readable text, furthermore, has a strict linearity. It fundamentally consists of one long concatenation of characters and spaces. In contrast to the physical page, plain texts cannot encode a difference between text and peritext via positioning. If peritextual aspects such as running titles, page numbers, and section headings are considered irrelevant, these need to be removed cautiously.

As was explained above, digital scholarly editions may be produced which can facilitate access to all extant witnesses of the text. It may be useful from a scholarly perspective to have a complete overview of the genesis of texts, but such a multiplicity of textual variants can complicate computer-based textual analyses. Many of the operations in the context of literary informatics are based on counts of

[322] Peter Shillingsburg, *From Gutenberg to Google: Electronic Representations of Literary Texts*, p. 21.

[323] See, for instance, Martin Volk, Lenz Furrer & Rico Sennrich, "Strategies for Reducting and Correcting OCR Errors", in: Caroline Sporleder, Kalliope Zervanou, & Antal van den Bosch (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Berlin: Springer 2011.

the occurrences of words, and if two different variants of one single text are both processed, this may obviously distort the statistics and overrepresent the frequencies of certain words. Scholars should therefore consider the texts that they include in their studies very carefully, and select the single variant which, in their view, best represents the text. Interestingly, it can be seen that, in this way, the use of algorithms promotes a return to a more idealistic view on textuality. The digital medium initially caused a decrease in the importance of an ideal and authentic version of the text, as limitations with respect to space in the printed editions did no longer apply.[324] While, in a digital edition, it becomes feasible to give expression to the view on editing as an open-ended discussion, the inevitable duplications that occur in such maximal editions clearly pose difficulties for scholars who aim to analyse the texts algorithmically.

Plain texts cannot reproduce aspects of typography such as font family, font size and colours, but they generally do include spaces, punctuation marks, capitals and line breaks. Such features of the text can be used, to some extent, to recognise the logical structure of the text. Whereas text mining has been described in Chapter 1 as a field which extracts information from unstructured data, it can be observed that texts generally do contain a degree of structure. Text mining may therefore be described more aptly as a field which exploits the formal and explicit clues which are present in the text to convert an implicit or inconsistent structure into an explicit machine-processable information structure. In most cases, text miming algorithms can only detect the presence of distinct logical units, nevertheless, and they cannot identify the nature or the function of these units. In most cases, including information about the function, the nature or the rhetorical effect of textual units demands a transition from data that represents explicit aspects to data that are descriptive and interpretative.

The lexical codes of a literary work can alternatively be represented via images of the printed page, in formats such as TIFF, JPG or PDF. A crucial disadvantage of these formats, however, is that the characters of the text cannot be manipulated until they are converted into machine readable text via OCR. Images can also reproduce the text's typography, which is a second textual aspect whose presence is manifest rather than latent. Images primarily provide a convenient access to digital versions of the original works, and they enable scholars to investigate the typographical aspects of the scanned pages on an individual basis. The development of algorithms for the automated analysis of the typography of large volumes of pages seems more challenging. Scholars who are interested in the

[324] Vanhoutte notes that the "electronic paradigm in scholarly editing has almost exclusively focused on the advantages of the size and economics of available storage capacity", and adds that the "digital archive as expanded text has in some cases jostled the one text away in favour of the multitude of many texts". Edward Vanhoutte, "Every Reader His Own Bibliographer - an Absurdity?", p. 109. In an important sense, the technical possibility to provide access to a large number of witnesses erodes the necessity of having to designate any single text as the most authoritative expression.

effects of typography generally concentrate on aspects such as indentations, font types and font sizes. At present, such characteristics cannot easily be extracted via digital image processing.[325]

4.5. Annotations

Computers can only process data which are present in an explicit form, or which can be derived consistently and unambiguously from other aspects which are explicit. Texts in natural languages, and works of artistic creation in particular, often have characteristics that complicate systematic querying. One important difficulty is that many of the characteristics that are of relevance to scholars are implicit.[326] An additional difficulty which hinders analysis is that natural language texts generally contain homonyms and synonyms. One concept may be referred to via distinct terms, and at the same time, one particular word may also refer to many different concepts. As the computer demands "complete explicitness and absolute consistency",[327] the impetus to study cultural objects via the digital medium implies the necessity to create digital surrogates in which all the properties that are implicit and imprecise in the original have been given an explicit and unambiguous expression. Smith explains similarly that, while the focus of the computer is on the literal characters that constitute the texts by default, the width of analytic procedures can be expanded if scholars encode "physical as well as semantic characteristics ... into symbol sequences parallel to the textual sequence". Smith suggests that such supplementary categories may be viewed as "strata that are parallel to and 'above' the textual sequence".[328]

---

[325] Lev Manovich has developed a method for the automated analysis of the style used in comic books by extracting data about "contrast, ... of texture and fine details, number of lines and their curvature". (Lev Manovich, "How to Compare One Million Images?", in: *Understanding Digital Humanities*, New York: Palgrave Macmillan 2012, p. 262.). These aspects do not seem relevant for the study of developments in typographical design, however.

[326] Among other aspects, plain text lack explicit data about the logical structure of the text. Because of this absence, it is impossible to distinguish the characters that appear in running titles or in the title pages from the actual body text of the literary work. In an acrimonious critique of the digital humanities, Adam Kirsch has argued that this inability to distinguish text from peritext has resulted in clear cases of unsound reasoning in studies based on big data. He illustrates his claim using a study conducted by Erez Aiden and Jean-Baptiste Michel. On the basis of the observation that references to specific years are most common in books published in that same year, the researchers claim that there is a general decline in historical awareness. Kirsch notes that this finding can be explained through the simple fact that the convention to print the year of publication on the copyright pages of books became more and more common. See Adam Kirsch, "Technology Is Taking Over English Departments: The False Promise of the Digital Humanities", in: *New Republic*, :May 2 (2014), n.pag.

[327] Willard McCarty, "Modeling: A Study in Word and Meaning", in: Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*, Blackwell, p. 258.

[328] John B. Smith, "Computer Criticism", 19–20.

Since the original sources do not include unequivocal markers of all the aspects that may be studied, such labels, if they are needed, must be supplied by the scholar. Implicit aspects can be made explicit via a process which, in this thesis, is referred to generically as "annotation". Structured annotations can be recorded via XML-based encoding. Stephen Ramsay notes about XML encoding that it can serve as "an elaborate scaffolding by which the vagaries of continuity can be flattened and consistently recorded".[329] The textual fragments to be annotated must initially be identified. Annotations may target words, phrases, sentences or texts in their entirety. XML offers a mechanism whereby fragments can be identified by placing a start tag and an end tag around the fragment to be annotated. The delineated fragments can subsequently be described by supplying a descriptive term either as the name of an XML element, or as the value of an XML attribute. Such a descriptive value can characterise the fragment as an instance of a particular phenomenon.

Within the humanities, the most widely used XML-based encoding language is the Text Encoding Initiative (TEI).[330] The standard was developed by a consortium of scholars in the humanities and the social sciences, and it currently consists of more than 500 descriptive terms.[331] Using the TEI, scholars can explicitly describe structural and semantic components of texts, such as paragraphs, sentences, place names, personal names and book titles. The TEI is a flexible and modular standard and was developed to support various forms of textual research. The standard provides facilities for "texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content". [332] The development of the TEI represents a major effort to standardise descriptive practices in the context of textual scholarship.

The TEI is an instance of an embedded mark up technique. The descriptive codes are interspersed with the lexical codes of the texts, and annotations can be distinguished from the text itself as a result of specific notational conventions. Annotations can alternatively be captured using data formats in which the descriptive values are separated from the plain text. Such external annotations can be recorded, for instance, using an entity-relational model that is implemented in a relational database. In this particular class of data formats, the descriptive values are contained in the cells of the various tables. The column names provide explicit information about the properties that are being described. External annotations can also be stored as statements based on the Resource Description Framework

[329] Stephen Ramsey, "Algorithmic Criticism", p. 8.

[330] James Cummings, "The Text Encoding Initiative and the Study of Literature", in: *Blackwell Companion to Digital Literary Studies*, Oxford: Blackwell 2007.

[331] An XPath query of the tei_all.xsd TEI schema which counts the use of <xs:element> returns a total number of 517 occurrences.

[332] "P5: Guidelines for Electronic Text Encoding and Interchange", <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html> (27 January 2014).

(RDF). RDF is a technique that can be used to formulate generic statements. It does not provide a vocabulary for such statements in itself, but it offers a general framework or a data model for making statements, which centrally consists of a structure containing a 'subject', a 'predicate' and an 'object'. According to Lee, this tripartite structure can be used to "say anything about anything".[333] The Open Annotation Collaboration (OAC) and the conceptualisation that was developed for the Nanopublications schema are two RDF-based techniques for the creation of annotations.[334] In the case of RDF-based statements, the object of the statement generally contains a descriptive value for particular text fragments, and the predicate often clarifies the aspects which are being described.

Embedded mark up schemes such as TEI can delineate text fragments via the placement of opening and closing tags, but systems for the recording of external annotations often lack a standardised method for the identification of the target of the annotation. Specific annotations about, for instance, the date of creation or about the overall theme apply to the work as a whole, and, in such cases, a reference may be included to the text in its entirety. When a finer level of granularity is required, however, data formats for the creation of external annotations can be combined with standards for the addition of embedded mark up. In the TEI, the passages that are outlined via the mark up can be identified via an @id attribute, and those passages can consequently be referenced in records in relational databases or in RDF-based statements.

Data formats for the storage of annotation can be used to describe aspects of the text which are latent, such as syntactic categories, phonetic transcriptions, lemmatised forms and literary devices. XML-based annotations can also be used to describe aspects which are wholly or partly explicit in the text, such as the typography or the logical structure. Mark up which concentrates on the typographical aspects of texts is referred to by Coombs et al. as "procedural mark up".[335] Such procedural mark up has been used in natively digital sources encoded in HTML or in EPUB. Mark up tags can alternatively identify particular units such as verse lines, paragraphs, titles or block quotes, without specifying the typographical appearance of these units. Next to the intrinsic characteristics that were enumerated in section 4.3, annotations may also capture extraneous or contextual data about the texts. For studies in the field of literary history, it is often necessary to collect data on the creation or on the publication of the text. Aspects such as the author, the title, the imprint and the general subjects can often be retrieved from library catalogues. Such supplementary data about the creation and the publication of texts can be referred to as metadata or as 'bibliographical data'. Additionally, in

---

[333] Lee Lacy, *OWL: Representing Information Using the Web Ontology Language* (Victoria BC Canada: Trafford 2005), p. 75.

[334] These two techniques will be discussed in more detail in Chapter 6.

[335] James H. Coombs, Allen H. Renear & Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing", in: *Communications of the ACM*, 30:11 (1987).

the case of online sources, it is also possible to collect data about the usage of particular texts. The server software that manages the access to a site normally keeps track of various aspects of the clients that request content, such as their geographic location, and the search terms that were used to find the text. Such data about the usage of documents can form a valuable source of information for research on the reception of literary works.

Structured annotations of the type that has been discussed can capture data which are supplementary, in the sense that they are not present in a manifest way within the original source. As the transformation of implicit aspects into explicit data almost inherently demands interpretation, the annotations that are furnished by individual scholars frequently have a subjective character. Different scholars may produce different observations about identical fragments. Johanna Drucker stresses that humanistic data are critically co-dependent on the observer, and that they are subsequently marked by idiosyncrasy and by ambiguity. Because of the constructivist nature of observations in the humanities, Drucker also objects to the term data, whose etymology suggests the existence of phenomena which are given a priori. Rather than viewing data as objective and pre-existing, they should be conceptualised as *capta*, or as observations which are taken.[336] Terms such as *facta* or *constructa* would perhaps be even more apt to emphasise the researcher-specific and the inventive nature of structured annotations.

It must be stressed, nevertheless, that this characterisation of data as subjective constructions cannot be applied to all the data within the field of literary informatics. In this chapter, a distinction was made between data which model observable properties of primary sources on the one hand and data which convert implicit and ambiguous aspects into structured and explicit annotations on the other. The aim of the former type of data is generally to objectively represent aspects of bibliographic items, such as printed editions or original manuscripts. A particular edition, once published, becomes a source which literary scholars can treat as a given. Plain texts and images may reasonably be viewed as objective data rather than as observer-dependent data. Additionally, while annotations that describe implicit aspects are often interpretative in nature, scholars may frequently arrive at a degree of intersubjectivity with respect to the aspects that are party or wholly explicit, such as the typography or of the logical structure. In the case of digitised texts, the correctness of the encoding that focuses on the logical structure of the text can be verified, to some extent, by inspecting occurrences of titles, paragraphs, line breaks or verse lines in the original. This verifiability confers a degree of objectivity on this type of encoding. Descriptive encoding which concentrates on the logical structure can never be fully objective or uncontested,

[336] Johanna Drucker, "Humanities Approaches to Graphical Display", in: *Digital Humanities Quarterly*, 005:1 (2011), par. 3.

however, as human interpretation will still be needed to decode typographical devices into the structural components they clarify.

Annotations can be created in two ways. Human scholars can firstly choose to make some of their understanding of texts available manually. When critics provide data by hand on a case-by-case basis, this means that the heuristics for the recognition of specific features do not have to be programmed into software applications. The manual creation of secondary data results in what Christoph Schöch refers to as "smart data". Schöch argues that the manual creation of digital data does not differ dramatically from traditional work in the humanities, as it mostly demands a meticulous and labour-intensive close reading of texts. Because of the dependence on manual work, smart data "does not scale well".[337] Alternatively, annotations can also be produced by making use of text mining. More concretely, this entails the application of algorithms which have been designed to detect specific items of interest within the text. The output of these algorithms can be captured as annotations.

Whereas data formats such as TEI, RDF and formats created in relational databases can have different technical properties, they simultaneously share two important characteristics. A first shared characteristic of annotations is that they are typically based on explicit ontologies. According to Davis et al., the creation of a model always implies a set of "ontological commitments", as it demands "a set of decisions about how and what to see in the world".[338] A surrogate is based on decisions concerning the aspects that must be included and the aspects that can be ignored. In the case of plain texts and images, this ontology is implicit. There is no formal document which defines the aspects of the source that are represented. The ontologies that underlie such objective representations can be reconstructed by carefully comparing properties of the model with the properties of the original. Structured annotations, by contrast, serve as explicit manifestations of latent textual phenomena, and are often created to allow for systematic processing. The textual phenomena that are studied can only be retrieved reliably if all the instances of such phenomena are consistently marked as such, using a fixed ontology. The conceptualisations that underlie TEI documents, relational databases or RDF statements are often available explicitly, in the form of XML schemas, ERD diagrams or OWL-based ontologies. Such explicit ontology files dictate the phenomena that may be observed within a particular domain, and additionally stipulate the vocabulary that may be used to describe these phenomena. Explicit ontologies most convincingly exemplify John Sowa's explanation that an ontology may be viewed as "a catalog of the types of things that are assumed to exist in a

[337] Chrostoph Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", in: *Journal of Digital Humanities*, 2:3 (2013), n.pag.
[338] R. Davis, H. Shrobe & P. Szolovits, "What Is a Knowledge Representation?".

domain of interest *D* from the perspective of a person who uses the language *L* for the purpose of talking about *D*".[339]

Structured annotations, secondly, are discrete in nature. Contrary to the words in linear texts, the meaning of individual annotations is not determined by the context in which they appear. A specific data value can usually be isolated completely from values that precede or follow that item. The division which is developed in this chapter between plain texts on the one hand and structured annotations on the other can be connected to the distinction which Lev Manovich discusses, in *The Language of New Media*, between narratives and databases. Manovich argues that novels and cinema have firstly "privileged narrative as the key form of cultural expression of the modern age",[340] while computers and the internet have introduced an alternative, non-linear mode of organisation, namely the database. Manovich presented narratives and databases as "natural enemies", as each of them "claims an exclusive right to make meaning out of the world". Rather than viewing database and narrative as two competing forms of expression, however, it seems more productive to regard narratives and databases as two distinct ways of organising information, each developed for a specific purpose. Narrative is the preferred format for human readers, while the database format is mostly needed to allow computers to process data in a systematic and in an efficient manner. In literary informatics, scholars generally convert linear linguistic compositions with a discursive or narrative structure into a database, which is essentially a collection of discontinuous properties and values, which collectively describe particular aspects of the original linear structure.

4.6. Derived data

In the previous section, a broad distinction was introduced between captured data on the one hand and structured and consistent annotations about these texts on the other. Structured annotations consist of explicit labels which may be connected to text fragments or to texts in their entirety, and which can resolve ambiguities in unprocessed plain texts. Both types of data mostly form a means to an end. Bates clarifies that distinct data values may be organised into larger constellations. Aggregations of such values may result in patterns in which "the sum of the elements constitutes something new, a whole with its own distinct qualities".[341] Such analyses are mostly performed to expose noteworthy characteristics of a corpus as a whole, or to identify individual texts with a conspicuously low or a conspicuously high value for a specific metric. Following Kitchin, the resources that

[339] John F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks / Cole 1999), p. 492.
[340] Lev Manovich, *The Language of New Media* (Cambridg, Mass.: The MIT Press 2002), p. 218.
[341] Marcia J. Bates, "Information and Knowledge: An Evolutionary Framework for Information Science", n.pag.

result from further manipulations of captured data or of structured data can be referred to as "derived data".[342] In this study, a distinction is made between two forms of derived data: (1) summations and (2) processed data. Summation, first, is a basic operation, consisting simply of a count of all the units which can be identified within the captured data or within the structured annotations. Word frequencies form an important example of such derived data. Summations can then be subjected to various forms of statistical learning techniques, such as clustering, counting or filtering. The numbers, lists or patterns that result from statistical analyses are described using the term "processed data".

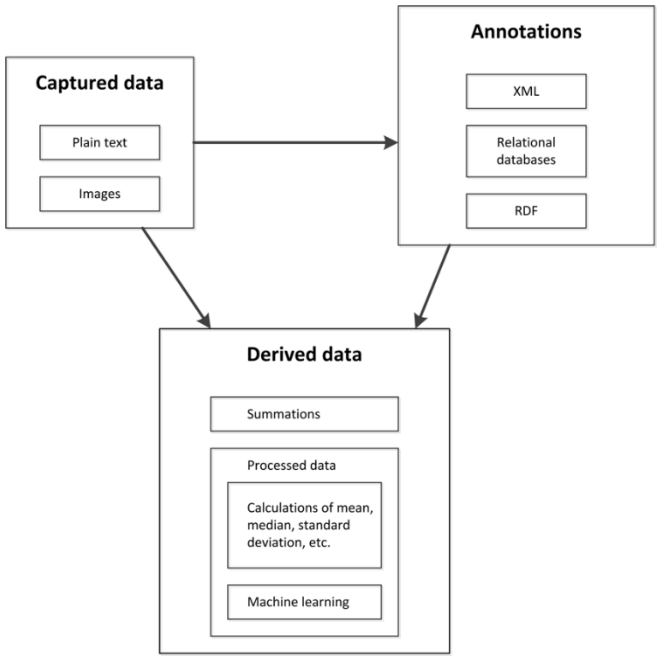

Figure 4.3. Three types research data, which can be analysed or created by three types of algorithms.

In this chapter, three types of research data have been demarcated. Three attendant types of algorithms may be distinguished. A first class of algorithms operates on unprocessed plain texts or images and aims to extract structured annotations which explicitly and consistently describe textual features which are ambiguous or implicit in the original works. This broad activity can be referred to as

[342] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*, pp. 7–8.

"data creation". A second class of algorithms takes these structured annotations as a starting point, and manipulates these statistically in order to produce derived data. This stage in the research process can be termed "data analysis". Thirdly, it is also possible to perform statistical analyses on captured data directly. In such cases, a single, more extensive algorithm combines the two tasks that can be clearly separated when statistical analyses are based on explicit secondary data. This third type of processing similarly begins with a data creation process. The results of the tokenisation process are stored temporarily, however, in a format that is secluded from the scholar who performs the text mining operation. Once individual textual units have been identified, they can also be analysed quantitatively. The fact that the results of the initial data creation process are not available separately can in some cases be a disadvantage. As procedures for the creation of structured data are often flawed, it is desirable to offer scholars the possibility to manually correct errors in the data that were generated algorithmically. Such interventions can improve the results of text mining processes. If the stages of data creation and data analysis follow one another immediately, there is a risk that analyses are based on erroneous data.

4.7. Conclusion

It has frequently been claimed that scholarship and science are currently being transformed by dint of the seemingly continuous advances in storage capacities and in database technologies. It is probable that the copious availability of machine-readable texts will also affect the methodology and the epistemology of literary studies. Literary informatics entails a form of research in which texts are not necessarily read fully by human scholars, and in which this task can be relegated, partly or wholly, to the computer. In a first phase of the research, digital surrogates need to be acquired of the literary works that are to be studied. At present, literary scholars have access to vast quantities of such digital surrogates, mostly as a result of the mass-digitisation programmes. Such projects normally model the observable aspects of original works via the creation of images or plain texts. In this chapter, such surrogates have been referred to as "captured data". Plain texts retain the inconsistencies and the linear nature of primary sources. Structured annotations, by contrast, result from a rigorous transformation of these resources, in which all aspects that are investigated have been classified unequivocally. These enrichments are necessary because of "the tension between the fierce formalism of code and the inexactitude of human practices and of natural language".[343] Structured annotations aim to transform capricious and ephemeral phenomena into tangible data values which can be processed systematically. they

[343] Caroline Basset, "Canonicalism and the Computational Turn", in: David Berry (ed.), *Understanding Digital Humaniities*, Basingstoke: Palgrave Macmillan 2012, p. 120.

are indispensable in studies which are based centrally on analyses of such phenomena. Statistical analyses of captured data and of annotations result in derived data.

The two broad phases that were distinguished within literary informatics - data creation and data analysis – can be clarified further using the concept of scholarly primitives which was first discussed by John Unsworth, and which has since been elaborated by a number of other scholars. Unsworth uses the term to refer to the "basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation".[344] They entail the basic forms of interaction with primary sources which result in an initial set of ideas about these objects. Unsworth argues that "discovery", "annotating", "comparing", "referring" and "sampling", "illustrating" and "representing" form the crucial scholarly primitives. Using Unsworth's concept as inspiration, Palmer et al. have similarly proposed an enumeration of the central activities that are common across academic disciplines. Palmer et al. make a distinction between core scholarly activities such as "searching", "reading", "writing" and "collaborating" on the one hand, and the more specific and discrete scholarly primitives which support these core activities on the other. The latter level includes basic acts such as "chaining", "browsing", "scanning" and "rereading".[345] Blanke and Hedges suggest a list of primitives which is more cognate to Unsworth's original explanation of the term. The authors claim that "discovery", "comparison", "delivery" and "collecting" form the cardinal activities within humanistic research.[346]

It may be claimed that the methods that are employed in literary informatics primarily provide support for annotation, comparison and discovery. Data creation, as discussed in this chapter, basically implies the process of creating annotations. Annotation refers to a process in which objects or fragments within objects are associated with particular descriptive texts. Applications which supply POS tags, lemmas or phonetic transcriptions enrich the bare tokens which are found in the original text with explicit and standardised values which allow for more systematic analyses, and such enrichments can, for this reason, be understood as annotations.[347] Data analysis, second, consists of a description of the differences and the similarities between two or more objects, and can, for this reason, be linked conceptually to the primitive which Unsworth refers to as comparison. This

[344] John Unsworth, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?".

[345] Carole L. Palmer, Lauren C. Teffeau & Carrie M. Pirmann, *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. (2009).

[346] Tobias Blanke & Mark Hedges, "Scholarly Primitives: Building Institutional Infrastructure for Humanities E-Science", in: *Future Generation Computer Systems*, 29:2 (February 2013).

[347] Annotation is based, in turn, on selection, which, according to Unsworth, entails both the identification of objects of interest and the identification of regions or fragments of interest within these objects. An annotation consists of a descriptive term which targets a particular text fragment.

core analytic function consists, more concretely, of the exposition of distributions, correlations or clusters. A distribution graph, for instance, enables scholars to compare the occurrences of particular literary devices in different stages of an author's literary career. Grouping and clustering operations such as k-means clustering or PCA similarly clarify the formal differences between the texts within a corpus.[348] While the results of comparisons may incidentally reveal relevant aspects in themselves, they can also form the bedrock of the scholarly primitive which Unsworth refers to as discovery. This third primitive encompasses the fortuitous identification of a document or of a text fragment whose distinctive or conspicuous characteristics warrant a closer examination. Unsworth stresses that discovery generally has a serendipitous aspect, as the process helps us to locate texts that can "become important to our work in ways that we would not have predicted, and therefore could not have sought".[349]

The discussion of the various types of data may help to clarify the nature of the term "big data". Many discussions of the nature of data in literary informatics reserve the term "big data" exclusively for collections of plain texts. In his discussion of "smart data" and "big data", Christoph Schöch maintains that the latter term refers to machine-readable texts produced by OCR software in digitisation projects, while the former term denotes data which have been created manually. In Schöch's view, smart data are prototypically represented by digital scholarly editions produced on the basis of TEI.[350] Julia Flanders and Matthew Jockers, in a discussion of the conflict between analyses on a micro-level and a macro-level, assume similarly that large-scale analyses of corpora typically take place on the basis of plain machine-readable texts, and that secondary data about the various phenomena that coalesce beyond the lexical codes can only be studied in smaller collections of manually encoded texts.[351] These readings of the term "big data" imply that data expatiating aspects that are implicit in the primary texts can only be small, and that big data are necessarily unstructured. This thesis offers an alternative view, however. In the case study that is conducted as part of this study, it has been shown that data about the linguistic and literary aspects of texts can be supplied both by human encoders and via text mining applications. When the

[348] Unsworth uses the somewhat antiquated term "sampling" to refer to the "result of selection according to a criterion". The process includes "the ability to show distribution and clustering". While Unworth does not explicitly discuss the differences between comparison and sampling, it may be argued that sampling implies a more specific form of comparison. The fragments which are selected through the process of sampling enables scholars to describe differences and similarities between these fragments. See John Unsworth, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?", n.pag.

[349] Ibid., n.pag.

[350] Chrostoph Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", n.pag.

[351] Matthew Jockers & Julia Flanders, *A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013*, pp. 5–7.

phenomena under investigation can be detected algorithmically, it also becomes possible to create big collections of smart data.