Some thoughts on appraisal in the digital age

von Ruud Yap

Digitalisierung stellt die Archivwissenschaft vor Herausforderungen. Diese Probleme zeigen sich auch in anderen wissenschaftlichen Disziplinen, die sich mit dem Auftreten von digitalen Objekten, Techniken und Methoden konfrontiert sehen. Diese anderen Fachrichtungen liefern vielleicht Lösungen, die für die Archivwissenschaft brauchbar sind. Ich habe untersucht, welche Möglichkeiten Big Data Technologie dem Archivar bieten kann. Indem wir Bedeutung quantifizieren, können wir das Zugänglichmachen und das Bewerten von Archivalien möglicherweise automatisieren. Indem wir Daten über unseren Gebrauch von Archivalien nach dem Google-Prinzip auswerten, öffnen wir neue Türen für die Bewertung; die Auffassung von Eric Ketelaar, dass jeder Umgang mit Schriftgut dieses anreichert, muss wörtlich genommen werden. Hinzu kommt, dass wir digitale Information in bisher unbekanntem Maße teilen können. So können wir unsere Herausforderungen mit den Nutzern teilen und gemeinsam mit ihnen lösen. Dadurch wird deutlich, dass der Satz von David Weinberger: „Die Lösung für den Überfluss von Information ist mehr Information" Gold wert ist.

Here I would like to explore the changes that have been brought about in archival appraisal theory due to the digitization of our world and thus the appearance of the digital record. Foremost I would like to propose some new views on handling appraisal and its pragmatic counterpart selection. I believe that most of our current approaches are still founded on pragmatic choices that were valid at the time they were made, but have no place in our current digital data centric world.

The qualities of digital records

Records first and foremost are logical entities. This we learn from viewing the archival record within the post custodial paradigm. Within its predecessor, the custodial paradigm, the record was still a physical entity under control of an archivist, which needed to be managed by an archivist. The custodial view started shifting in the twentieth century when archivists were confronted with the exponential growth of information. This growth demanded a faster processing method and archivist found themselves positioned not at the end but at the start of the information food chain.

The life-cycle model in which records were eventually transferred to an archive no longer was adequate: archivists should be involved at the moment of creation of records. The Australian archivist Peter J. Scott recognized this when he was confronted with an ever growing flow of information in the 1960's. In order to handle this growth of information he formulated the 'series system'. In this system he separated the record in context and (physical) object. The focal point became the context of a record: controlling the context of creation, management and use enlarged the view of the archivist. He or she had to look beyond the 'fond', beyond the records that were under his direct control.[1]

1 Barbara Reed, The Australian context relationship (CRS or series) system: An appreciation, in: Cunningham, Adrian ed., The Arrangement and Description of Archives amid Technological Change. Essays and Reflections by and about Peter J. Scott (Brisbane 2010) 347.

Scott however was still working with paper records and his ideas became into full growth by the introduction of digital records. Digital records exist due to a specific combination of hard- and software. Preserving this specific combination can be problematic. The digital record is dynamic: it changes; it has to change in order to be preserved. It is therefore more important to determine which result of the combination, we consider to be important. A digital record can have multiple manifestations over time (different formats, visualizations) but will still be the same logical record. Meaning, or rather the fixation of meaning, will have to take place outside the physical record. Not the object itself, but the context, captured in metadata, now determines the meaning, authenticity and value of the record. With metadata we build a cocoon of meaning. It helps us to capture context and at the same time provides us with the tools to determine and share meaning. Records therefore always consist, and this is not new, out of objects and metadata.[2]

The relation context and object is best described via the records continuum model: 'Archival documents first and foremost provide evidence of the transactions of which they are a part – from this they derive their meanings and informational value. The function which produced the record defines its meaning. Meaning is therefore dynamic, because the continuum-based approach suggests integrated time- space dimensions. Records are 'fixed' in time and space from the moment of their creation, but recordkeeping regimes carry them forward and enable their use for multiple purposes by delivering them to people living in different times and spaces.'[3] Different functions, different meanings. More importantly this states that records can have different meanings at the same time, depending on the recordkeeping regime (context) in which the record is used. Digital records thus do not hold a single physical form, have a single meaning or a single use. Geoffrey Yeo expressed this best by adopting the sociological concept of the boundary object introduced by Susan Leigh Star and James Grasmere. The boundary object is an object which is shared between separated communities. These communities use different contexts in which they use the object. The boundary object serves as an interface between these communities. This concept demands a reevaluation of the archival appraisal process.[4]

Why do we appraise?

If context is everything, then what is appraisal? The formal answer would be that, according to Peter Horsman, appraisal is setting criteria to determine if and which records should be retained.[5] Appraising is making choices: a legal, administrative and historical valuation. More specifically, according to Hans Waalwijk, it is ascribing value to documents at different times and in different processes.[6] It is followed by selection: the administrative processing of appraisal decisions.[7] In the day to day operation, however, appraisal has a pragmatic consequence: by determining which records to retain, we know which records can be disposed of.

The periodic disposal of records is a necessity, although it was not always common good between archivists. Although he was not the first, it was Theodore Roosevelt Schellenberg who started the 'modern' thinking on appraisal, selection and disposal in 1971 by acknowledging that records could have different values for different users by defining a primary and secondary value of records: 'the primary value of a record, according to Schellenberg's definition, is evidentiary: does the record show evidence of action? The secondary value, nearly equally as important, is for research: would a researcher want to use this record to understand more about the context of its creation?'[8] This roughly meant that if records had lost their primary value and did not have a secondary value they no longer needed to be retained. At the beginning of the 21st century the process of appraisal also shifted from object to context. Records became 'value laden instruments' and it became important to understand the record and its context.[9] Joan M. Schwartz and Terry Cook proclaimed that archives were social constructs with which power could be exercised and power could be denied. Power in hands of the record creators, users and archivists: 'Like archives collectively, the individual document is not just a bearer of historical content, but also a reflection of the needs and desires of its creator, the purpose(s) for its creation, the audience(s) viewing the record, the broader legal, technical, organizational, social, and cultural-intellectual contexts in which the creator and audience operated and in which the document is made meaningful, and the initial intervention and on-going mediation of archivists.'[10] Appraisal, selection, retention and disposal became instruments of power. Archiving was not just a mechanical act of retaining records, but also a social construction of reality.[11] Appraisal should therefore concern the context of a record.

Earlier I stated that the appraisal, selection and disposal of records are necessities. First of all there are several legal obligations to dispose and destroy records. Privacy regula-

2 Sue McKemmish, Glenda Acland, Nigel Ward and Barbara Reed (1999), Describing Records in Context in the Continuum: the Australian Recordkeeping Metagegevens Schema (Geraadpleegd op: 11–11–2010, http://infotech.monash.edu/research/groups/rcrg/publications/archiv01.html).
3 Sue McKemmish, Yesterday, Today and Tomorrow: A Continuum of Responsibility, in: Proceedings of the Records Management Association of Australia 14th National Convention, 15–17 Sept 1997, RMAA Perth 1997.
4 Geoffrey Yeo, Concepts of Record (1): Evidence, Information, and Persistent Representations, in: The American Archivist 70 (z. p. 2007).
5 Peter Horsman, Abuysen ende Desordiën. Archiefvorming en archivering in Dordrecht 1200–1920 (Amsterdam 2009).
6 Hans Waalwijk, Een bouwsteen voor de toren van Babel. Over definities voor waardering, selectie en verwijdering, in: Brood, Paul ed., Selectie. Waardering, selectie en acquisitie van archieven, Jaarboek Stichting Archiefpublicaties 2004 (Den Haag 2005).
7 Horsman, Abuysen ende Desordiën, 36.
8 John Ridener, From Polders to Postmodernism. A Concise History of Archival Theory (Duluth 2009) 84.
9 Ibidem 124–125.
10 Joan M. Schwartz and Terry Cook, Archives, Records and Power, in: Archival Science (z. p. 2002) 4.
11 Ibidem 5.

tions for instance state that certain categories of records are to be destroyed after a certain period of time. One can debate the necessity of these regulations and even their motives, but these regulations exist and as such determine the faith of some records.

Theoretically as I mentioned earlier, the processes of appraisal and selection contribute to the value and meaning of a record. Appraisal is a necessary condition for the creation of records. I find that a concept derived from dynamic semantic sciences describes the value of appraisal best. Incremental interpretation is used for the interpretation of sentences. According to this mechanism new information is constantly added to old information and the interpretation of a sentence in a text is affected by sentences that precede or follow that sentence. A context can be seen as an information state and the meaning of a sentence as a function that changes these states of information.[12] The concept of 'possible worlds' is used: a mechanism to model information to consider a certain state of information as a collection of possibilities. By adding new information one eliminates possibilities. The context of that sentence determines the possible meanings of a sentence and by adding more context (more sentences) a single meaning is achieved.[13] This mechanism can also be described as a form of updating.[14] Appraisal can be described as updating. Appraisal ads context and thus updates the record and determines its value and meaning. Similar to Eric Ketelaar and Derrida whom state 'that every interpretation of the archive is enrichment, an extension of the archive. That is why the archive is never closed. It opens out of the future. The archive, in Derrida's thinking, is not just a sheltering of the past: it is an anticipation of the future.'[15] The mechanism of incremental interpretation however is reversed: the 'semantic incremental interpretation of sentences eliminates possibilities, while appraisal opens the records to new possibilities and meanings. Appraisal is thus essential for the creation of records: by adding context (value) we transform information into records and logically dispose of information that is not worthy to retain as a record.

There is also a biological motivation for the disposal of records. Douwe Draaisma a professor in the history of psychology describes the functional necessity of forgetting as such: a human being needs to forget in order to function normally. Remembering everything comes at a great cost. He describes the therapy of Eye Movement Desensitation and Reprocessing (EMDR). This therapy is used to treat traumatized patients in which they seem to forget (or dispose) of some aspects of the trauma. The context of the trauma is updated and meaning is added while the object (trauma) is separated from certain contextual elements (the traumatic context). This approach seems very successful for patients to coop with their trauma. Another example were forgetting seems necessary for the functioning of the human mind was illustrated by Viktor Mayer-Schönberger in his book Delete in which he recounts the story of an certain AJ who possessed a 'hyperthymestic memory': a nearly perfect memory. She could barely function in normal life because got stuck in the past. The perfect memory made it impossible for her to make any decision because she could take every earlier situation into account.[16]

Appraisal selection and eventually disposal of records have a similar effect: they remove contextual elements from the archive and by doing so they update the context and add meaning to the archive. Appraisal also helps us to identify and take into account useful and relevant information in order to decide, reconstruct, learn and act.

Despite the earlier mentioned necessities for the disposal of records there is a tendency to ignore these necessities when dealing with digital records. Digital records are stored and our storage capabilities seem endless. This probably is correct. The mechanical act of archiving digital records, storing files, is not problematic. Everyone here probably owns several terabytes of information. On a business level we store exabytes of information. We can all buy storage at a ridiculous discount. On a more scientific level Gordon Moore of Intel already claimed in 1965 that the costs of storage would keep declining due to miniaturization.[17] The storage of information will never be the problem and the mechanical act of archiving in the sense of 'storing' is solved.

Despite this 'solution' there remains a functional necessity to the disposal of digital records: the costs of keeping records either digital or on paper are high. The costs of preserving and understanding the information we store, will be and in some cases already is a problem.

Because storage itself is combination of hard- and software. The costs may be declining but longevity is far from guaranteed. Hardware and software change fast, according to the standard Moreq2010 organizations replace their technology every 3 years. This means that a digital record that is retained over a period of 75 years will be managed by more than 25 different combinations of hard- and software.[18] The best strategy would be to achieve system interoperability which would ensure that the mechanical process of migration is lossless (without information loss)[19] This demands strict requirements for describing, storing and exchanging information. Metadata, format control,

12 Martin Stokhof en Jeroen Groenendijk, Betekenis in Beweging, in: Algemeen Nederlands Tijdschrift voor Wijsbegeerte Volume 90 (1998), 26–53 (staff.science.uva.nl/~stokhof/bib.pdf, acces date: 11–11–2010).

13 Ibidem 3.

14 Henriëtte de Swart, Introduction to natural language semantics, Center for the Study of Language and Information – Lecture Notes 80 (Stanford 1998) 130.

15 Eric Ketelaar, Tacit Narratives: The meaning of archives, in: Archival Science 1 (z.p. 2001), 138.

16 Viktor Mayer-Schönberger, Delete. The Virtue of Forgetting in the Digital Age (Princeton 2009) 21–22.

17 Source: http://nl.wikipedia.org/wiki/Wet_van_Moore (Accessed on: 4–3–2012).

18 DLM Forum Foundation, MoReq2010 ®: Modular Requirements for Records Systems – Volume 1: Core Services & Plug-in Modules, 2011, published online (http://moreq2010.eu/) 23–24.

19 DLM Forum Foundation, MoReq2010 ®: Modular Requirements for Records Systems – Volume 1: Core Services & Plug-in Modules, 2011, published online (http://moreq2010.eu/) 151.

etc sought after by initiatives in open data and linked data. A more pragmatic, but nevertheless valid approach is controlling your flow of information by appraisal, selection and disposal. This not only mitigates the risks of information loss, but also addresses the more conceptual needs for appraisal.

Appraisal and selection in the digital age?

Fact is that control seems nearly impossible. There is a deluge of information. Digital information is stored without prejudice. We retain information unconsciously. We store terabytes of information. But the digital deluge is a reality. So let's entertain the idea that all digital information we create is retained. This information needs to be accessed and understood and we, archivists, are responsible for organizing this. How can we, archivists, tackle this problem?

I turned to the world of data-intensive science. In this world the collection, retention and use of large volumes of data and information is common practice. Data-intensive science or the more popular term big data science even has its own paradigm in Informatics: research based on large volumes of digital data with the help of digital technologies has been named the Fourth paradigm. Form, volume and function have changed some dramatically that a paradigm change occurred.[20]

Data-intensive science seems to face the same challenges as us archivists: the exchange of information, the mutual understanding of information is a challenge. The datasets within the different scientific disciplines are not interoperable: not only on a technological level, but also on a cultural and linguistic level.[21] Janus seems to rule big data as well. Access and readability are also problematic for big data science: 'the vast amounts of data have greatly reduced the value of an individual data element, and we are no longer data-limited but insight-limited.'[22] The solution to this problem is appraisal and selection: filtering the information so that only relevant information and data is presented. The complexity of 'everything' only leads to doubt. Even science does not need a hyperthymestic memory. Disposal of data or information however isn't an option for big data science. They opt for a forgetting without forgetting by applying filters. It are these filters, or rather the instruments that apply the filters that I find interesting for archival appraisal application.

One of the tools in use is the so called workflow tool:'a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks, acting like a sophisticated script'[23]. The benefits working with workflows is that all the steps in a workflow can be automated and every step is controllable because it is documented.[24] The archivist notion of the genre can be applied in such an instrument. Genre can be defined as the description of the arrangement of the formal characteristics and content of a document, but a broader definition is provided by language and communication sciences in which 'genre' is described as a representation of a communicative action. Genre here describes not only form, but also purpose, participants, timing and location of the action.[25] Our experience in describing the context of records can be put to use in formulating automated workflows: we can filter records by determining which forms, purposes (or functions), participants, timing and locations have value and 'asking' the workflow to select those records which meet certain desired combinations of these elements on the basis of their metadata. It even allows us to appraise and select records upon creation: if a certain combination of form, purpose (or function), participant, timing and location arises at the moment of creation a record can automatically be appraised and selected by a system. It would allow us to automate appraisal.

Big data also recognizes the potential of all data and offers us ideas how to recombine and reuse information. We all expulse enormous amounts of data in our day to day business: every time we create, open, use, change, search, find or copy information, these acts are monitored and in most instances logged. Our data expulsion, a term coined by Viktor Mayer-Schönberger in his book *Big Data: A Revolution*, is a treasury of information. We should utilize this information by considering all interaction with records as valuable. The notion that every interpretation of the archive is enrichment, an extension of the archive as Ketelaar and Derrida stated must be taken literally.[26] Google certainly adopted this view by applying using all data that is created by users in their products. Google translate works because it uses all input (even misspellings) to predict a fitting translation. Relevance is predicted through the use of clicks on search results combined with the time a user stays on the website behind the result, the number of scrolls, downloads, etc. Every interaction is analyzed and used.

Archivists can also use this approach to appraise the vast amounts of information that is retained and stored. At the National Archives of the Netherlands we have used this approach to appraise our own collection. We acquired funding to digitize 10 % of our collection. It was therefore vital to determine what that 10 % should be. In short which records should be digitized first. We used the interaction of

20 Clifford Lynch, Jim Gray's Fourth Paradigm and the reconstruction of the Scientific Record, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 177.
21 Mark R. Abbott, 'A new path of science?', in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 115; Carole Goble and David de Roure, The Impact of Workflow Tools on Data-centric Research, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 137.
22 Abbott, A new path in science?, 114; Charles Hansen, Chris R. Johnson, Valerio Pascucci and Claudio T. Silva, Visualisation for Data-Intensive Science, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 162.
23 Goble, The Impact of Workflow tools, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 138.
24 Ibidem 142.
25 T. Yoshioka, G. Herman, J. Yates and W. Orlikowski, Genre taxonomy: a knowledge repository of communicative actions, in: ACM Trans Inf Syst 194 (z. p. 2001) 431–456, 433.
26 Ketelaar, Tacit Narratives, 138.

our patrons with the archive, number of recalls, to determine which records were deemed valuable and were digitization would have the most impact. This was just a small recombination application of data. On a larger scale we can analyze the interaction of record creators with digital records. All document management, content management database applications log and store (meta)data concerning our interaction with the records retained within that system. We could reuse and recombine that metadata to determine what correlations exist between those interactions and the value of the records. Those correlations could help us appraise information automated and on a large scale. The number of searches and changes done by a civil servant could tell us something about the meaning of a record. The number of communications or signatures could tell us something about the value of a record. There are no guarantees, but the current approaches which still rely on substantial human intervention and handling are not tenable in an information- and data centered world.

The fourth paradigm also tells us to embrace the bits and bytes. In some ways it tells us to still use the 'physical' qualities of the digital record. Digitization has made our information machine readable: information can be processed by machines. Yunhyong Kim and Seamus Ross researched the automation of the genre classification. They tried to quantify the genre indicators such as form, purpose (or function), participant, timing and location.[27] Kim and Ross experimented by transforming stylistic and visual features of genres into classifiers and adding statistical models to those classifiers. They, for example, compared documents word counts and pixel values of images of individual records to define certain genres of records. The results were positive: automated appraisal on the basis of their classifiers approached the results of an 'average human labeler'[28] The scope of this research was limited, but it shows that automated selection is a viable option and with better classifiers can be compared with the quality of human appraisal. This is hopeful because automating appraisal and selection has several benefits: large volumes can be handled in less working hours and at greater depth than humans could ever hope to do. An automated approach could also solve the problem of granularity by which I mean the low level of granularity of archives. Our ability to describe, appraise and select records has been limited by the human factor. The money, people and time to make archives accessible at the level of the record rarely were available. The level of detail of the common finding aid or inventory is low. For example: an average record description at the National Archives of the Netherlands translates to 10 cm of paper. The aggregation of description is high due to limited resources. These principles are easily translated to the process of appraisal and selection were we utilize high aggregation levels. Putting machines to work will reduces the human factor and will open up our archives at a more detailed level.

I firmly believe these approaches can help us appraising and selecting our records. There is a vast amount of information that can be explored and used to aid us in our tasks. I must state however that technology alone can never be the solution. As I recounted before archiving is not just the mechanical act of storing information. Similarly automating our appraisal process by the use of digital instruments will not replace the human factor.[29] 'It has also been suggested that massive data mining, and its attendant ability to tease out and predict trends, could ultimately replace more traditional components of the scientific method. This viewpoint, however, confuses the goals of fundamental theory and phenomenological modelling. Science aims to produce far more than a simple mechanical prediction of correlations.'[30] Technology helps us to filter information faster and maybe better, but value and meaning is ultimately added by the human intervention and interaction. Big data enables us to automate appraisal and selection. It broadens the use of records because we can reuse and recombine information to open up the archive even more.

Don't do it yourself

In the Netherlands the committee on Appraisal and Selection, led by professor Charles Jeurgens, concluded that integral archiving would be too costly due to the human effort. Keeping those archives accessible and understandable would be too costly. The economic motive is an impure, but valid argument. As I have said, no technological solution will reduce the costly human factor.

Digital records do provide us with an opportunity to ease the burden of the archivist. I already discussed that digital records do not hold a single physical form, a single meaning and a single use. I can easily add that digital records are not bound to single physical location and can be accessed remotely. We could open up digital archives radically and enable our patrons to participate in the archive.

In 2008 the Finnish archivist Isto Huvila came to the conclusion that participation is the way ahead. He formulated three characteristics that defined his participatory approach[31]:

1. Decentralized custody:
   Custody over the archive is shared between archivists, record managers and participants in the archive who collectively share knowledge on the records, their context and their use;

27 Gillian Oliver, Yunhyong Kim and Seamus Ross, Documentary genre and digital recordkeeping: red herring or a way forward?, in: Archival Science 8 (z. p. 2008) 296.
28 Ibidem 56.
29 Peter Fox and James Hendler, Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 150.
30 Paul Ginsparg, Text in a Data-centric world, in: Tony Hey, Stewart Tansley and Kristin Tolle, The fourth paradigm: data-intensive scientific discovery (Redmond 2009) 190.
31 Isto Huvila, Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management, in: Archival Science 8 (2008) 15–36.

2. Total user orientation:
   Usability and retrievability have the highest priority. Those principles guide appraisal and retention. The archive is oriented and reoriented on its users.
3. Contextualization of records and the archival process:
   Context is not only derived from provenance. Interaction adds value to the records and the archive.

We could allow users to participate in the process analyzing the results of our new and automated processes of appraisal and selection. Their interaction will add value and meaning to the records. Their interaction will add to our analysis and workflows. Sharing digital records and sharing digital information will not be difficult. It all depends on our willingness to embrace the qualities of digital records and the challenges they pose and to use them to our advantage. The digital deluge must not be countered by building larger dikes, but by learning how to keep afloat on the digital ocean.

Conclusion

So to conclude I would like to leave the reader with these ideas:

- Digital records have other qualities than paper records: use those differences!
- The digital deluge demands appraisal and selection: everything is nothing
- 'The solution to the overabundance of information is more information' (David Weinberger – Everything Is Miscellaneous)
- Open and share your archives

Thank you, try the veal. ■

Ruud Yap
Nationaal Archief, Den Haag
ruud.yap@nationaalarchief.nl