INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2–3, 2009, 129–138

Digital Discoveries in Museums, Libraries, and Archives: Computer Science Meets Cultural Heritage

Antal van den Bosch and Jaap van den Herik
Tilburg centre for Creative Computing, Tilburg University, The Netherlands

Paul Doorenbosch
Koninklijke Bibliotheek, National Library of the Netherlands, The Hague, The Netherlands

A harmonious combination?

The question that underlies this special issue is: can the arts, humanities, and sciences (in the Anglo-American sense) exist in harmony? Their perspectives on nature and culture are so different that it is not obvious that they would converge if brought together, and that the result would be harmonious. Yet, we see computer science approaching the arts and humanities and the other way around. To understand the mutual attraction, we start by describing their idiosyncratic behaviours.

When asked to tell about the computer on his[1] desk, the typical computer science researcher will answer that the machine is a particular instantiation of a universal Turing machine (Turing 1936), capable of doing an infinite amount of things with data and information. The same machine will be described by a curator of cultural heritage data as a useful storage and data-accessing device that is helping to save precious time. Let us assume that the two meet to discuss collaboration. The computer scientist will not be surprised to see the utility of the device, as storage and access are two of its basic strengths. With half suppressed impatience, he will inquire whether the curator has considered moving beyond merely digitizing, storing, and accessing data. What about accessing and discovering information and knowledge? The curator will respond by pointing at the advanced state of metadata standards in the cultural heritage world. He may point at the Dublin Core Metadata Initiative, for instance. Upon browsing the Dublin Core specifications, the computer scientist may spot key phrases such as *Resource Description Framework*, and will be duly impressed.

Yet, the computer scientist is quick to point out that one of the weaker links in the digital era of cultural heritage remains the human user, struggling with

© Institute of Materials, Minerals and Mining 2009
Published by Maney on behalf of the Institute
DOI 10.1179/174327909X441063

the extra orders of magnitude of data that he is expected to handle as it becomes available in digital form. Then, computer scientist and curator both face the prospect of dealing with centuries of hand-coded metadata. The computer scientist discovers inconsistencies, missing elements, and mixed taxonomies. The curator becomes a little uneasy, and admits that although searching basic data around 2010 is easier and faster than ever, adding metadata remains the preserve of the cultural heritage expert, who is limited by the attention span, the working hours, and all other limiting features of the average human being. Having reached this point in the discussion, the computer scientist walks to the whiteboard and begins drawing blocks and diagrams that must lead to a personal curator assistant of the future.

Thus, in a caricature, this is the starting point of convergence of a growing amount of interdisciplinary work between computer scientists and cultural heritage curators exemplified in this issue of *ISR*. All around the world, state-of-the-art computer science is, and is soon to be applied to new challenges in the access and use of cultural heritage. Here we highlight a particular research programme that can be seen as representative of the new domain of interdisciplinary collaboration. The Continuous Access to Cultural Heritage (CATCH) programme is funded by the Netherlands Organisation for Scientific Research (NWO). More precisely, it is a coordinated effort from the Dutch cultural heritage institutions together with two NWO divisions, Physical Sciences and Humanities.

In nine contributions from teams operating within the CATCH programme, the issue highlights such diverse topics as automated metadata enrichment, handwriting retrieval, cross-collection search, and personalized museum tour generation. The preface serves as their introduction. Before we summarize some of the key lessons learnt in CATCH so far, we turn our attention to the present and future of CATCH as a whole.

Continuous access to cultural heritage

Since 2005, CATCH has funded research teams that focus on improving the cross-fertilization between scientific research and cultural heritage. Each team consists of a PhD student, a post-doc researcher, and a scientific programmer. To ensure transferability and interoperability, the research teams carry out their research at the heritage institutions, according to the *laboratorium extra muros* formula. Currently, CATCH is financing 10 research projects conducted in nine cultural heritage institutions. Recently, CATCH has received additional support from the Ministry of Education and Scientific Research to fund four more projects.

Looking back on four years of CATCH, its impact on the Dutch cultural heritage sector can be said to be significant. Here we refrain from a project-by-project analysis but provide a telling instance of a project that occurred at the national library of the Netherlands (in Dutch, the *Koninklijke Bibliotheek*, henceforth KB). The KB was one of the founding fathers of CATCH, and took part in designing the overall proposal. After an incubation time of two years

in which cultural heritage and computer science learned to understand each other's language and way of looking at the same object, the CATCH programme was written in a relatively short period and in close harmony. Initially only a limited number of people in the R&D (research and development) department of the KB saw the potential of the programme. However, the KB director at that time was instrumental in voicing the opinion that CATCH was necessary to bring a new kind of expertise into the library, to be able to keep up with the rapid changes in our information society.

This was a visionary opinion, since libraries are currently going through significant changes, which holds true for their staff as well. Next to people trained as librarians, a growing number of IT specialists are being hired. So, we see library issues nowadays framed as computer science challenges, as libraries (as well as museums and archives) have to deal with increasing volumes of digital data, metadata and the Web. The big advantage of the development is that through the potential of the digital information environment, cultural heritage institutions have more opportunities to attain their primary goal: to provide the best possible interaction between (1) users and (2) objects, information and knowledge.

In the first two years after its launch, the CATCH project affiliated with the KB worked in relative isolation, despite efforts to organize opportunities for exchanging experiences. Yet, after this period interaction began to happen. Random personal contacts at the coffee machine were an important catalyst (the research team did spent at least three days a week on a regular basis inside the library premises). By presenting specific library questions to members of the research team and asking them their opinion, a mutual understanding started to emerge.

Of course, this contact was wished for, but the CATCH programme designers had also anticipated these developments by emphasizing the connection between science and daily practice. They deliberately planned two roles in every project. The first role was to be played by a cultural heritage institution employee, aware of the institute's processes, and having the ability to participate in the scientific discussion. The other role was given to a scientific programmer, who was given the task to build software prototypes to show how scientific results and scientific inputs could be used in the primary process of the institution. The formula worked out very well, not only in the library environment such as that of the KB, but also in museums and archives, which were at that time even less attuned to advanced IT.

In retrospect, we may remark that in 2004 it was unlikely that an archivist would guess that a supercomputer would ever be used to 'google' 17th century handwritten material. At that time, a musicologist could only dream that there might be algorithms capable of retrieving large amounts of songs stored as audio or in music notation form. Also, in those days the director of a museum might believe that an excellent website builder was a real asset to make an appropriate visitor interface. Now all directors are convinced of the added value of an academic approach.

In 2005, six projects started their research; two years later, another four were selected by competition. Together with the cultural heritage institutions

involved, these 10 projects reached their conclusion in 2008, and their results were so promising that they wanted to transform the prototypes they had developed into full-scale applications. This led to an offspring of CATCH, the recently started implementation and validation project called CATCHplus.

Moreover, the successes in CATCH did not escape notice by the Dutch Ministry of Education and Scientific Research. So, in 2009, they commissioned NWO to organize a new round of competition for acquiring a CATCH project. Four projects were awarded. In the selection procedure, the CATCH organization was supported by the ISAB, the International Scientific Advisory Board, the members of which are renowned scientists from all over the world.[2]

All in all, by participating in CATCH, the cultural heritage sector was able to raise interest in disclosure and access issues in a digital environment, and find support for it in a field of science with which cultural heritage practitioners had hardly been aware. Awareness of new methods and different ways of approaching traditional objects and knowledge has clearly increased throughout the sector. The remaining question is: what will be the future of this harmonious combination? We can only speculate, but we do offer our view on one type of challenge.

On 12 March 2009, the breaking news in the cultural heritage world was the solution of the *Nachtwacht puzzle*, which had lasted for 367 years. In 1642, the famous Dutch painter Rembrandt van Rijn finished his masterpiece entitled *De Nachtwacht* ("The Night Watch"), in which 21 persons are depicted. From the outset there existed a list of names for those depicted in the painting; even the amount of money paid by the people for being included in the painting was known. Yet, for one or another reason there was no written account of which name matched which person. Two well-known figures were known — Banning Cock and Willem van Ruytenburch — depicted at the very forefront of the painting. Because they were identified, their names have been passed from generation to generation in the education of all Dutch youngsters.

Meanwhile, historians were curious to solve the riddle of the remaining names and figures in Rembrandt's creation. This *Who's Who* exercise turned out to be a real challenge for museums, libraries, and archives. The Dutch Historian Bas Dudok van Heel was finally able to solve the puzzle adequately by very accurate research. He brought many things to light (Dudok van Heel, 2006; Van Raaij and Van Zeil 2009). We single out a few of them: (1) the name of Banning Cock should be Banninck Cock; (2) Jan Clasen Leijendeckers passed away in 1640, two years before the painting was completed; (3) Jacob Jorisz, the drummer, is not on the name list as he did not pay 100 florins (he earned 40 florins a year).

Taking this puzzle as an example, assume that the information from all museums, libraries, and archives was available for online perusal. The intriguing question then is whether a computer program could be given the task of assigning the proper names to each of the figures painted by Rembrandt. Such is the prototypical CATCH challenge of the future. First, it involves the

wide range of seriously complex computer science challenges: a massive cross-collection search and analysis, leading to the representation and aggregation of all analysed information into a wide web of knowledge, concluded by an inference process working on this web. Second, it challenges collection managers and researchers to be the crucial human part of the loop of a next-level organization of their domain's knowledge.

Discoveries made and lessons learnt

During the exercise that has now been under way for four years, researchers in CATCH are looking back on a range of discoveries, from the hoped for and the expected to the rather unexpected. In this preface, we will not dwell for very long on the expected results; some of the issue's contributions provide excellent examples. The application of advanced yet existing computer science methods to cultural heritage data almost instantly led to practical findings. One such was that collection databases with errors (i.e., virtually any reasonably sized collection database) can be cleaned much more quickly if the human expert is aided by an error-detecting computer program (Van den Bosch *et al*., 2009). But welcome as such results are, they do not teach us anything fundamentally new.

Most of the unexpected discoveries in fact involve significant human aspects because cultural heritage is fundamentally a human endeavour. When it meets technology, even if based on scientific principles, the domain experts and collection managers react according to their prime concern: to keep the original data and objects safe from harm — and this extends to metadata as well. An important first step in every CATCH project, and we believe in most successful interdisciplinary undertakings that bring together the cultural heritage with computer science, is establishing confidence in all participants that the effect of cooperation will be additive, never destructive.

It is relatively easy to reassure a collection manager of the effects of improved access of digital metadata. Reassurance becomes harder when computer science methods work to enrich data and metadata automatically, for instance by suggesting the addition or correction of metadata. By providing suggestions, the computer does not harm the data, but it does enter the human realm of expert knowledge. The initial reaction of many cultural heritage researchers and collection managers is one of disbelief. How could a computer make sensible suggestions, when it has provably not gone through the motions of becoming what they themselves are? The technical answer to that question may be hard to accept: under certain conditions, computers can infer from previous knowledge or examples how an expert would classify or analyse cultural heritage objects. The reassurance here is that the computer is not taking over from the experts, but is there to help them in their task. More precisely, and reassuringly: the human in the loop remains essential for the computer to operate.

Summarizing what CATCH has brought to light, we highlight below three types of encounter between computer science and the cultural heritage sector.

The subsequent contributions in this special issue, which we summarize in the next section, provide more details and examples of these encounters.

1. **The shock of scaling up from cases to databases.** Many databases in the cultural heritage field have been painstakingly compiled by manual data entry in the course of years or decades. Each case (record) in the database may have been constructed with great care, over a long time, and with the help of many other resources. The shock comes when, after these cases have been put into a database, the computer inspects the complete collection in milliseconds or less and comes to over a thousand conclusions almost instantly. Errors are thus revealed, new metadata indicated and certain items recommended as possibly relevant to the expert. Such speed and comprehensiveness are simply not possible for a human being. To witness the effects can in practice be a great shock. Yet, when the expert understands that the computer has indeed done a passable to good job on thousands of cases in the blink of an eye, he quickly begins to see the potential for optimizing his own workflow. Thus, the time and care invested in individual objects and cases could be improved. Moreover, the load of certain other data management jobs (such as eliminating errors from the data) could be alleviated by the computer's suggestions.
2. **Formats and technologies may change by the season.** Many curating practices are centuries old and have proven their durability through time. This cannot be said of computer technology. In this respect, the scepticism of cultural heritage curators is understandable. Up to now, most computer hardware technologies have become obsolete within ten to twenty years. Durable digital data storage is therefore a continuing challenge. The situation with computer software technologies is better, but only mildly so. Due to the fact that hardware and software technologies are market-driven, the future can be expected to remain changing in uncertain ways. The only realistic approach to the future of digital cultural heritage is therefore to take into account this uncertainty. Any plan involving digitization, further processing of information, and enrichment of cultural data must be made robust against future changes. Bad experiences from the past (losses of data, of inaccessibility of data in old formats, irreplaceability of old computer hardware) are abundantly available to learn from.
3. **New possibilities require new criteria.** Computer science offers new possibilities that often have a new dimension. For instance, computers scale well in terms of storage, retrieval, and computation. The quantities of data that can be handled by computers are orders of magnitudes larger than a human can process, or that a physical depot or archive can handle. Without the old physical limitations, cultural heritage institutions now ponder the question whether they actually want to provide access to all those objects that were inaccessible before. There

used to be reasons for a museum to have certain objects on display, and others stored in the depot. Now, the question must be asked anew: could and should all digitized objects be made accessible?

This issue: an overview

This special issue aims to offer a cross-section of state-of-the-art computer science solutions to issues in accessing cultural heritage (including archaeology and natural history). Summarized into just a few key phrases, the contributions of this special issue are about annotation, retrieval, and personalization. The media accessed and annotated cover a broad spectrum: handwriting, text, music, paintings, archaeological objects, photographs of objects, speech, and radio and TV broadcasts.

Some topical threads run through the contribution. Rather than summarizing the contributions one by one in their actual order, this overview groups them by the most salient research threads: annotation, retrieval, and personalization. We remark that one thread that is not exclusive to any subset of research contributions, namely metadata, is an overall thread that binds all areas. Importantly, metadata is to a large extent the medium that helped start the interdisciplinary collaboration in the various projects, as metadata is a well-understood concept in both scientific communities. The three other threads are invariably tied to aspects of metadata.

Annotation is the best represented thread in this issue. In most of the contributions that deal with annotation, the task involves the automated or computer-assisted enrichment of a heritage object with metadata. Luit Gazendam, Véronique Malaisé, Annemieke de Jong, Christian Wartena, Hennie Brugman, and Guus Schreiber describe and evaluate a system that generates suggestions for metadata annotation in their contribution entitled *Automatic annotation suggestions for audiovisual archives: Evaluation aspects*. Gazendam and colleagues discuss the issues that arise when part of a cognitively demanding task is left to computers. Does the computer offer sufficient quality? When it suggests metadata that a cataloguer would not assign to an object, is the computer's suggestion wrong?

In *Digital support for archaeology*, Paul Boon, Guus Lange, Laurens van der Maaten, Hans Paijmans, and Eric Postma showcase a number of solutions of metadata assignment to archaeological objects, and further enrichment of existing textual metadata such as field logbooks. Interestingly, Boon and colleagues discovered that even human experts find certain categorization tasks very hard to perform, and have trouble explaining how they perform them. Added to the fact that annotated and categorized archaeological object databases are typically small, computers cannot simply be programmed or trained to mimic the experts. Instead, the team discovered that the computer could be helpful when it was used to provide visualizations of an entire collection of objects in a single image, clustering all objects according to mutual similarities in visual features such as contours, shape, and texture.

Sometimes metadata are not a high-level abstraction of a limited or fixed number of object attributes but stay close to describing the object data, only abstracting so much in order to be better searchable or understandable. In this issue, we find two such studies, on music and handwriting, respectively. In *Modelling folksong melodies*, Frans Wiering, Louis Grijp, Remco Veltkamp, Jörg Garbers, Anja Volk, and Peter van Kranenburg provide an in-depth overview of existing and new approaches to modelling and retrieving music. Working with a collection of Dutch ballads recorded in the first half of the 20th century, the authors aim at discovering similarities between these ballads, in order to provide new insights into the mechanisms of oral transmission — the only way ballads were passed on before the radio and gramophone era. The project combines this goal with providing a musical search engine. The goals mutually strengthen each other, as high-quality ballad retrieval (finding the most similar ballads to any single ballad) must make use of knowledge on how ballads are copied, changed, and mixed in oral transmission. Second, in *Where are the search engines for handwritten documents?*, Tijn van der Zant, Sveta Zinger, Lambert Schomaker, and Henny van Schie start by explaining that reliable writer-independent automatic handwriting recognition is still not possible. Yet, in particular constrained situations, the technique can work fairly reliably. With the human expert in the loop, the machine can learn from individual annotations of specific stretches of handwriting and find similar stretches of handwriting that signify the same letters and words in hundreds or thousands of other places, in digitized images of handwritten documents, at a scale that no human could physically perform.

As the final instance of the annotation thread, the contribution by Antal van den Bosch, Piroska Lendvai, Marian van der Meij, Marieke van Erp, Steve Hunt, and René Dekker, entitled *Weaving a new fabric of natural history*, focuses on letting computers suggest improvements to an existing metadata scheme. While the past decades have seen a surge in the development of digital object databases, only recently have the first international standards been formulated for blueprinting an object database. Hence, many existing databases need an upgrade. One way to automate this upgrade is to analyse automatically the conceptually weak but nonetheless often used 'comments' or 'miscellaneous' fields that serve as an unstructured collector of otherwise useful information, but that were not given a place in the outdated database design. Second, the study introduces a way to discover names for the relations between database fields. The study uses a natural history object database as its working example; for instance, the method discovers that some animal typically 'occurs in' a country.

Retrieval is the end goal of the aforementioned contributions by Wiering *et al*. and by Van der Zant *et al*., as search engines are usually thought to provide the most directly usable and tangible kind of interface to cultural heritage objects, such as pieces of recorded music, or images of handwritten documents. For the same reason, retrieval is also mentioned by the other projects involved in metadata annotation. Yet, with *Information retrieval in*

cultural heritage, by Marijn Koolen, Jaap Kamps, and Vincent de Keijzer, this issue has a contribution that focuses in particular on the special requirement that cultural heritage institutions have for search engines: that they offer unified access to their many heterogeneous data and metadata collections. Searching should, in principle, be possible not only in text, but also in textual metadata. Furthermore, the search engine should be intelligent in ranking and presenting heterogeneous best matches to a given query, and it should be sensitive to the different levels and registers of language used in data and metadata.

Beyond the relatively straightforward searching in text and textual metadata, searching for and retrieving audio and video broadcasts offer additional technological challenges that are addressed in *A multidisciplinary approach to unlocking television broadcast archives* by Laura Hollink, Bouke Huurnink, Michiel van Liempt, Johan Oomen, Annemieke de Jong, Maarten de Rijke, Guus Schreiber, and Arnold Smeulders. Apart from textual data (such as subtitles) and textual metadata, it is vital that multimedia search in a broadcast archive such as investigated by Hollink and colleagues genuinely exploits similarities in visual elements between video shots. Similar challenges are addressed by Willemijn Heeren, Laurens van der Werff, Franciska de Jong, Mies Langelaar, Roeland Ordelman, Thijs Verschoor, and Arjan van Hessen, in their contribution *Easy listening: Spoken document retrieval in CHORAL*. Their focus is on retrieval from spoken word collections, and their technological focus is on developing accurate automatic speech recognition software to create a reliable metadata layer of recognized words.

Personalization, the third thread, is at the heart of *Cultivating personalized museum tours online and on-site* by Yiwen Wang, Lora Aroyo, Natalia Stash, Rody Sambeek, Yuri Schuurmans, Guus Schreiber, and Peter Gorgels. Wang *et al*. aim at developing a new framework for enriching a person's museum experience through the use of computer science methods. A web-based tour planner is described, that interactively probes the visitor's preferences and interests, and generates a tour through a museum that best matches the visitor. The *tour wizard* can be used off-line and not in the museum, and may attract people to come to the museum; alternatively, the wizard can be used in a portable device to be carried through the museum in a live visit, enhancing the visitor's experience.

The special issue starts with the latter contribution; it then switches to the annotation thread, which fluidly merges into the retrieval thread.

Acknowledgements

The authors wish to thank the Netherlands Organisation for Scientific Research (NWO) and the Dutch Ministry for Education, Culture, and Science (OCW) for their sustained support and funding of the CATCH programme. Annemarie Bos and Annejet Meijler are recognized for their support at the start of CATCH. Mark Kas, Christien Bok, and Rosemarie van der Veen-Oei have been particularly instrumental in setting up the organizational framework on which CATCH has been able to develop.

Notes

[1] In this contribution, we use 'he' and 'his' whenever 'he or she' and 'his or her' are meant.

[2] http://www.nwo.nl/catch — Last visited March 2009.

Bibliography

Dudok van Heel, Sebastian. 2006. *De jonge Rembrandt onder tijdgenoten: Godsdienst en schilderkunst in Leiden en Amsterdam*. PhD diss., Radboud Universiteit, Nijmegen, The Netherlands.

Turing, Alan. 1936. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* ser. 2(42): 230–65.

Van den Bosch, Antal, Marieke van Erp and Caroline Sporleder. 2009. Making a clean sweep of cultural heritage. *IEEE Intelligent Systems* 24(2): 54–63.

Van Raaij, Ben and Wieteke van Zeil. 2009. Puzzel van de Nachtwacht na 367 jaar opgelost. *Volkskrant*, March 11, 2009.

Notes on Contributors

Correspondence to: Antal van den Bosch, Tilburg centre for Creative Computing, Faculty of Arts, room D343, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands.

Email: Antal.vdnBosch@uvt.nl