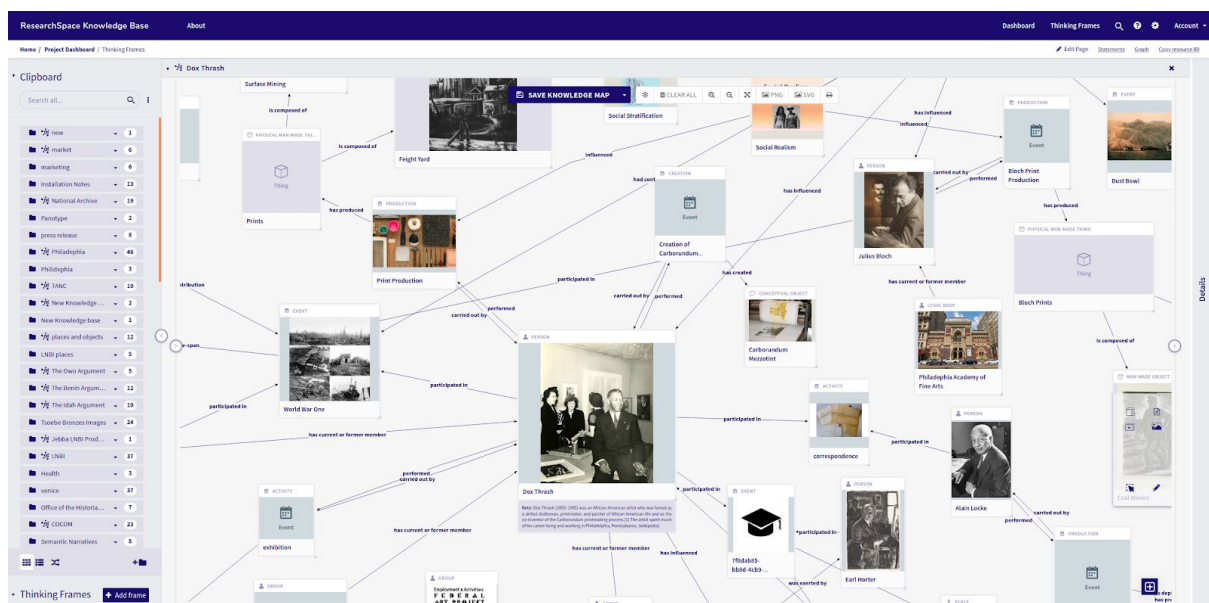# Collection Data and Social Justice
Breaking the cycle of Babbage in creating and sharing Collection Data

Dominic Oldman
British Museum
ResearchSpace - http://www.researchspace.org

Proceedings of, The Art Museum in the Digital Age - Jan. 2021

Slide 3

This presentation is about bias and issues of diversity in collection data and the contribution that collection systems and associated digital research systems have on these problems. There are many reasons why this bias occurs, despite the claim of neutrality. The Andrew W. Mellon Foundation has pointed out in their surveys, issues of demographics in terms of those employed in Cultural Heritage organisations, and in the artists and works that are selected and presented. Bias is also linked to issues of language, of breadth of expertise and experience, broader discrimination, and in the resources given to researching different parts of the collection.

However, this presentation has a focus on the 'form and structure' of the collection system (model) itself and how its structure contributes and perpetuates bias and lack of diversity. The narrowness of documentation, its static and dated nature, its lack of a two way dialogue, the absence of historical context, its failure to link across time, and crucially its authoritative stance which excludes other vantage points and other supportable 'knowledges'.

This is an aspect not commonly considered but has a considerable effect on the way in which cultural heritage is represented in a data form, significantly impacted by commercial software. The limitations and inherent narrowness of collection data systems is a crucial question since what we previously hid in internal systems, supporting internal processes, is now used for public communication exposing historical, gender, cultural and racial bias from the back office into the public domain and therefore it is not simply a technical problem, but an organisational one, which requires new approaches.

Slide 4

The tool illustrated allows for complexity. It is not deterministic and allows researchers and cultural heritage experts to assert contextualising information whether that be other historical events, ideas, arguments, influences and so on. This is one of a number of different tools which could only have been designed by cultural heritage experts themselves. They are designed to put the thinking and knowledge back into cultural heritage documentation and there are no differences with current systems in terms of the effectiveness and efficiency of processing. It provides better search and discovery by using a conceptual framework, with richer results and better digital preservation of knowledge that is independent of the software.

In fact it reduces overhead by reducing fragmentation caused by the gap that collection systems impose. The ability to include information that is historically relevant, to allow a collaboration of different viewpoints or vantage points, to combine the qualitative and quantitative, to explore the different parts or histories of an object in relation to other objects, technology, modes of production, periods, events, social relations, and concepts of the mind. It saves data with contextual relationships - it does not simply technically link URIs of legacy data. It provides a data narrative and a querying mechanism that better reflects the questions that people want to ask, and contribute to, across different heterogeneous datasets. We will come back to this.

Slide 5

The Philosopher Andrew Feenberg, in developing a theory of instrumentation shows how the process of knowledge generation is reduced by the constraints of conforming to instrumental and institutional frameworks. Computer systems, because they are based on processes determined by technologists and developers, are "underdetermined" by a fixation on efficiency and responsiveness that, particularly in the current generation of standard computers, is completely unnecessary. We can have representation that is far closer to the way we, as humans, abstract and understand information with no new practical technical issues. The problem is that commercial software houses have no understanding of how those abstractions work or how to support them - they can only pre-determine using commercial patterns of representation.

They are not subject experts so they provide systems that conform to their own knowledge systems and business models, not the knowledge systems of their users who simply accept this as a natural limitation of computers. The Cultural Heritage community is now more than capable of producing systems that do exactly what these software companies cannot do. Represent ideas, provide rich contextualised interconnections across disciplinary areas, compare and resolve issues of different 'knowledges' and allow computers to help us solve complex problems. These systems will not come from the software houses - they will come from the community when we have unattached ourselves from the tendency to attend to their needs, rather than our own.

Slide 6

I want to use history to illustrate how computer databases have impeded the progression of 'knowledges'. Different 'knowledges' are potentially part of a whole and in rejecting traditional Western approaches of scientific truth, parts can be combined to provide a better type of objectivity. Western bias in online data resources preserves old ways of looking at and describing concepts. Note that many of the people mentioned in the slides and with a data card are commonly referred to as **the Father's of**. This is a common term often referring to Western men. Francis Bacon himself, rejected the idea that knowledge stopped at the Pillars of Hercules (the rocks of Gibraltar) and that other knowledge existed and could be developed. The ordering and preservation of past knowledge was the preoccupation of the medieval era - **Static Knowledge.** I present the static data card in wikipedia to tell you almost nothing important about Francis Bacon, but Bacon was about expanding the framework for new discovery and new conceptual categories.

Slide 7

As the Western world developed ways of exploiting resources, including human ones (the first computers were human), Adam Smith, theorised about the generation of wealth and its relationship to value. His two major books. The Theory of Moral Sentiments and the Wealth of Nations, are two parts of an overall whole. While the division of labour was a device for potentially greater productivity and wealth production, which he believed would benefit all classes, rich and poor, Smith also understood that this could also have a detrimental effect on people and society - later categorised as alienation - both through the commodification of the labourer and the products that they produced. The product of collection data has the hallmarks of a production line that can alienate.

He believed that the these effects needed to be mitigated by a strong public domain and social framework - a sentiment that has some re-emergence today with issues of modern

slavery, environmental damage and global wealth inequality - issues that are embbeded in the cultural heritage objects that we curate. However, his disciples were quick to separate these different considerations - to remove the interconnection between them and claim they were no inherent dependencies.

Slide 8

The 'Father' of Modern Computing - Babbage was a disciple of one part of Smith's thesis. He wanted to make workers invisible. He saw how the use of human computers, originally skilled mathematicians solving questions of astronomy, could be further broken down such that the calculations could be performed by an army of untrained and unskilled workers. But more than this - it was then possible to produce a machine that could take over calculations completely. The Astronomy Society, (which included Babbage), was made up of business people, accountants and lawyers whose interest in astronomy was directly connected to business risk and international trade.

Slide 9

Standards are part of the problem when it comes to intellectual tasks. They provide the ability to claim neutrality and link information - but what they do is remove the need to think when describing data in computer information systems and are designed to benefit the production of reductive software. In an article by Langmead et al in the International Journal of Digital Art History, they say that collection data input has become known as 'grunt work'. To address issues of quality they suggest the role of Data Steward to stand over the 'grunt' workers, to ensure a reliable transfer of information from source to field. The scene starts to look like a factory line with the shop steward ensuring that the predetermined, one size fits all input forms - are processed to the factory owners satisfaction. How can this build dynamic knowledge? It is a division of labour that reduces the quality of the output in favour of a quantity with limited use value, but it never occurs that it can be any different - that is what technologists have told us. Data input is a non-intellectual task recording non-intellectual information and in doing so has inbuilt obsolescence. They have, with our assent, pre-programmed the knowledge out of the task.

Slide 10

As a humanist, the father of Science Fiction, was more interested in the wealth of education and human knowledge. The answer to many of the world's problems was the integration of knowledge from different parts of the world. Well's foresaw that technology would reduce the distance between people and between the knowledge they communicated to each other. This didn't mean imposing the knowledge of one part of the world onto the other, but to synthesise as a way of improving overall knowledge. It would require collaboration and understanding of other ways of knowing. Wikipedia's supposedly neutral stance continues to provide evidence for continuing research into Western bias and international differences.

Slide 11

n/a

Slide 12

During the computer revolution of the 20th century even management consultants realised that progression meant a new paradigm to achieve productivity. Instead of top down management hierarchies, businesses needed to decentralise and utilise the skills and judgement of knowledge workers, encourage creativity, promote continual learning and an entrepreneurial culture throughout the organisation. The type of task management or scientific management developed by Winslow Taylor (Taylorism) and which had been practiced by Babbage, would only impede productivity, stifle innovation and limit progression.

Despite Drucker's logical prediction of a knowledge society, businesses and organisations have been reluctant to loosen the reins and still protect their control of capital and costs - limiting the effectiveness of knowledge workers. Institutional computer systems are still designed to serve management objectives more than user's needs, despite the lost opportunities and the wasted time servicing systems that divert knowledge workers away from more productive pursuits. Artificial intelligence and machine learning are the digital equivalents of Babbage's mechanical automations. To Druker, computer's still had only one task, to provide intelligent humans with the raw information they needed to make intelligent decisions. Humans solved complex problems - and the computer's role - he called computers "morons" - was limited.

Slide 13

To people like Doug Engelbart, the father of the mouse and part of the teams that led to Advanced Research Projects Agency Network (ARPANET) providing the foundations for the modern internet, computers could do more than simply provide and filter information, they could help people perform increasingly complicated tasks including ones which involved complex decision making and digital collaborations. These researchers were not pure computer scientists, but psychologists. For example, they could help humans design complicated buildings taking into account more interconnected factors that had a bearing on creating the perfect spaces, they could manipulate symbols or they lay out complex arguments. With the ARPANET system computers could help solve more complex problems by connecting people and 'knowledges' and help them to utilize it. H.G. Wells had much more in common with Engelbart than with Wikipedia.

Slide 14

In the meantime the manufacturer of computers (by dominant companies like International Business machines - IBM) had different design visions based solely on the needs of business. The innovations of the world wars, aimed at breaking codes and weapon navigation systems, triggered a race to get the first commercial computers to a mass market. The focus was to solve business problems and businesses had no need for systems that dealt with the complexities of historical knowledge. Database systems were not designed to do comparative history but were designed to process consistent quantities of data quickly for businesses on their terms. Cultural heritage institutions adopted this model without question conforming to the already established, form and structure. They created single perspective reductive datasets. Libraries conformed the most because of their particular relationships with the publication industry, but attempts to homogenize museums have been limited and

localised, but overall highly damaging to modern digital strategies.

Slide 15

The first significant European cultural heritage digital project was RAMA, Remote Access to Museum archives. It set the tone for all subsequent projects. Regardless of the technology of subsequent projects the aim was the same, to homogenize data. Even in the age of the Semantic Web the same mindset and standards are applied. Ontologies don't reflect the world and its histories, they preserve an old order of artificial vocabularies designed to satisfy the requirements of technologists and costs, not subject experts. Despite the changes in technology mediums, whether client server, ETL, Service Oriented Architectures, Semantic Integration or data warehousing, the technologists are the modern day equivalents of the medievalists, seeking to preserve a way of thinking and preventing new ways of knowing.

In the 1990, one of the biggest names in cultural heritage information David Bearman, founder of the Museums and the Web Conference, after attempting his own version of RAMA in the United States, concluded that we were going in the wrong direction. Bearmans says that "*Increasingly it seems that we should have concerned ourselves with the relationships...between the objects.*" not lavish huge attention on standardised vocabularies. Large integration projects wanted to do it all, serve commercial services, educate, provide resources for research, engage the public, and despite RAMA, Acqualle, MESL, Project Bamboo, Europeana, and many others, all have had very limited success juggling these objectives.

Slide 16

RAMA partnered with CIMI, The Consortium for **Computer Interchange of Museum Information** (initiated by the Museum Computer Network Group. The idea was that they could archive with the same efficient results as the libraries. The MARC standard had provided a platform for software companies to create cheaper 'off the shelf' systems, without any regard for the intellectual aspects of libraries. Publishers and other commercial companies could also exchange book records.. But users were left with systems that had no understanding of the intellectual interconnections of the books themselves particularly across disciplines. In Schnapp and Battles, the Library Beyond the Book, they say - "The stack itself offers seemingly limitless opportunities for the prepared mind to find conjunctions and synchronicities or wander productively. And yet those possible conjunctions among the books, however vast in number, are limited - not only by the size of the library… but by the catalog, by whatever ordering principle determines which books stand next to another on the shelfs." Digital systems potentially provide, through books, an interconnected map of the history and ideas of the world, but instead we get a map which is, in their words, a "tattered ruin".

Slide 17

In the 1990s, an article at MIT by Howard Shrobe and Peter Szolovits, showed how computers could help with human reasoning, inference and collaboration using a common conceptual reference based on the real world which could be understood by humans and computers. By sharing a universal but expansive language that everyone understood, human users could explore more complex relationships between different aspects of the

world that the computer could follow - and point out new knowledge that was too complicated for the user to see easily themselves. It could even understand things that were not all uniform and belonged to different categories at the same time, - just like the real world. It could support fuzziness and more natural questions.

In the software development world developers started to dismiss real world ontologies. They said that they were getting in the way of putting raw information on the web, regardless of whether it was interpretable. In reality, it would have meant greater collaboration and a change in the way that developers worked, but they just wanted to get on with playing with a new technology. As a result Linked Data has failed to come anywhere near its potential and even Tim Berners Lee has started again on the Web of Data. Commercial collection system suppliers could simply stand back and watch. In reality some ontological frameworks were needed to work with Linked Data and developers who have engaged with it have simply transferred existing artificial vocabularies into an ontological form and have not engaged with knowledge representation or new thinking.

Slide 18 - not used
Many academics have taken issue with the reuse of commercial software, commercial practices and narrow standards. Notably Manfred Thaller in looking into the issue of how computers could be made useful to historians concluded that relational databases and technical approaches were not designed to address researcher's questions. For Thaller we needed to look at the specific needs of different researchers and understand how to design systems that reflected needs. The position of Historians was summarised in the 1970s when Charles Tilly, a pioneer of computer aided research, said that many historians "consider the accumulation of systematic knowledge about human behavior either impossible, dangerous, of little value, or a serious diversion from other more worthy ventures". Little has changed just as the underlying form and structures and/or the associated mindsets employed to process historical data has changed little whether relational databases or Linked Data..

Slide 19

The notion of conceptual reference was not an invention of the CIDOC CRM. It was a term used in Cultural Heritage to describe how curators, archivists and librarians combined the two aspects of their work. The physical aspects and the intellectual (or external contextual) aspects. For example, when we research an exhibition, we do not have a list of objects lined up in a row according to some arbitrary category. They are arranged to answer questions, provide a narrative, to make connections to ideas, places, and people outside the confines of the institutional walls. Publishing legacy catalogue data to the public works against this.

Slide 20

Collection, archival and library systems effectively broke the link between basic physical and administration arrangement and essential intellectual processes, as part of a general widening of the gap between the different directorates. However, the administrative side of cultural heritage organisations have realised that this division has created a problem as new calls emerge to rectify issues of diversity and bias. Back office processes in the digital age, like documentation, directly affect front office engagement. Organisations have sleep walked into a digital dilemma. They are knowledge organisations that have digital systems with significant legacy content problems that don't reflect or integrate internal knowledge, let alone support other 'knowledges'.

As such there are now many organisations that have initiated operational projects to remedy this problem, but they rely on the very same suppliers and commercial systems that harbor this inherent problem. This includes looking at the possibilities of Linked Data where exactly the same split has taken place between the technologists who promote technical Linked Data and the subject experts who assert the richer ontological commitments. As such Linked Data, again starting with the Libraries, has simply replicated this division. As such they are highly unlikely to make more than a ripple in the goal to "decolonialise" or historically contextualise their data. This can't be done without reuniting all parts of the conceptual reference and making it more inclusive and by supporting community software that breaks the stranglehold of Babbage's invention.

Demonstration

The following demonstrations are not based on code but on data relationships. The logic and meaning is embedded in the data in the way in which we would normally think about - and therefore we can start manipulating it in a way that is more to anyone working with it. The interfaces are driven by intelligent data.

What you are looking at is the CIDOC CRM framework in action.