Impact of Data Digitization

THIS REPORT IS SUBMITTED TO **DR. DORJE DAWA** IN THE FULFILLMENT OF MONTH-LONG PROJECT FOR THE FIELD OF HISTORY, CULTURE & CIVILIZATION AT B. TECH. VTH SEMESTER, CIC

Aayush Jain | B. Tech. Vth Sem | 11601
Aditya Sharma | B. Tech. Vth Sem | 11602
Vaibhav Jain | B. Tech. Vth Sem | 11634

ABSTRACT

It’s digitization of data, that has allowed us to make this report in collaboration, online and in real time while we’re comfortably sitting in our homes. Our reports, pictures, books, movies and art is on the internet. Oh and from ‘our’ we mean the beginning of times since humans started keeping records!

Be it biblical texts and historical images or memes and celebrity quotes, everything is stored online nowadays in an organized manner. There was a time, not so long ago, when libraries and video rentals were a thing. Now it’s kindle and Netflix and chill.

Many texts written on papyri which were illegible to human eye have now become much clearer through multispectral and hyperspectral imaging. Talking about the documents of the past, all of it is data waiting to be analysed. All of this is studied under digital humanities.

In this project we have tried to explore these changes that our methods of keeping records have undergone and how these have in turn impacted our ethics and emotions.

Table of contents

Introduction

Digital Humanities

It has only been a couple of years since we began hearing the term Digital Humanities being uttered quite prominently, though mostly in academic circles. For the uninitiated, it almost sounds like an oxymoron. After all, for most practical purposes the digital and humanities have always been seen almost as contradictory terms, existing in distinct silos. But as data proliferate and play an ever-growing role in our life decisions, the division between engineers who design algorithms and social science experts who interpret those data cannot hold anymore. The content industry via media, entertainment and culture, together with the public make an extensive use of Big Data and Machine Learning techniques. Thus comes the role of Digital Humanities.

The digital humanities, also known as humanities computing, is a field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities. It is methodological by nature and interdisciplinary in scope. It involves investigation, analysis, synthesis and presentation of information in electronic form. It studies how these media affect the disciplines in which they are used, and what these disciplines have to contribute to our knowledge of computing.

Digital humanities scholars use a variety of digital tools for their research, which may take place in an environment as small as a mobile device or as large as a virtual reality lab. There are thousands of digital humanities projects, ranging from small-scale ones with limited or no funding to large-scale ones with multi-year financial support. Some are continually updated while others may not be due to loss of support or interest, though they may still remain online in either a beta version or a finished form. The following are a few examples of the variety of projects in the field:

Cultural analytics

1. This refers to the use of computational method for exploration and analysis of large visual collections and also contemporary digital media.
2. Cultural analytics also includes using methods from media design and data visualization to create interactive visual interfaces for exploration of large visual collections e.g., Selfiecity and On Broadway.

Analysis of macroscopic trends in cultural change

1. Culturomics is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts.
2. Researchers data mine large digital archives to investigate cultural phenomena reflected in language and word usage.

Digital Heritage: Old and New Sources

Sometimes the discussions about the digital transition in the humanities turn into a polemic between interpretation of the Old versus the New Testament. Either you, as a humanities researcher, join the fashion of the new media, or you stay faithful to materialities such as paper sources or books, and restrict yourself consequently to established methods such as content analysis and source criticism. For one reason or another, the digitization of sources has been accompanied by raising a barrier between ‘old’ and ‘new’ sources and ditto researchers. What is the basis for this division and what does it say about the future of the humanities? The distinction between old (material) and new (digital) sources is not only theoretically nonsensical, it is even practically damaging for the academic profession.

The inseparable connection between technology and knowledge already existed long before Plato objected to the rise of writing as a replacement of memory. The replacement theory has since never disappeared; from writing to the typewriter and from the printing press to the computer: new inscription devices and storage technologies were invariably seen as a replacement of, or a threat to, the old. And while typewriters and printing presses are becoming virtually extinct in daily communication between people these days, it is nonsensical to argue that collections of writings, pictures, film tapes, and other non-digitized sources have become superfluous. The conclusion that older media or collections derived from ‘old media’ can be discarded because the ‘contents’ now exist in digitized form somewhere in the world is like saying that we no longer have to conserve paintings because of the invention of photography. After all, not just the materiality of the source matters, but also the indissoluble tie between materiality, production, and distribution technology, and the selection of sources at a particular historical moment.

The digitization of sources is not just a technological issue; it deals with production and curation of content. Still, the possibility of converting all kinds of sources into digital files triggers two contrary impulses: the one extreme is to want to store, from now on, every single digitized or digital born utterance; and the other extreme is to discard all ‘original’ sources once they have been digitized. Both extremes are implications of the replacement theory: the belief that we can record everything with computers as a result of which everything non-digital becomes superfluous. Whoever believes in the possibility of complete inscription and storage of every single utterance should learn from history that this idea has recurred over the past five centuries.

Need of Digitization?

Data Digitization is the process by which physical or manual records such as text, images, video, and audio are converted into digital forms.

Benefits

Digitized data offers the following benefits

- Preserving cultural assets with digital copies
- Data is much more accessible now
- Better follow-up of documents
- Simplified access to content
- Easier and quicker sharing of documents
- Limited search time thanks to a more accurate classification of data
- Reduced printing costs
- Increased space (23,000 pages fit in a compact hard drive)
- Reinforced security for storage and exchanges
- Digitalization of data has changed communication

Digitization in Action: Case Studies

Digitization has been a huge impact on our lives in many ways. Some of these can be observed easily around us. But perhaps, we sometimes fail to pay attention to many ways which might not influence us. Here we would like to discuss about some of the brilliant efforts done in the field.

Case Study: The Griphos project: reassembling ancient frescoes

The reconstruction of fragmented objects is of great interest in archaeology, where artefacts are often found in a fractured state. Vast quantities of fresco material, commonly excavated from archaeological sites in thousands of pieces, remain in storage indefinitely, unexamined and unstudied, as resources are simply not available to complete time-consuming manual searches for any fragments that match. Frescoes are usually of unknown design, with many pieces of the same colour, with few clear distinguishing features and often with missing sections. Fragments are generally fragile, and excessive handling is undesirable, at best, and highly damaging, at worst.

The Griphos Project (in Greek, griphos means ‘puzzle’ or ‘riddle’) was established to develop computational methods to aid in the reconstruction of fresco fragments. Griphos has been utilized, so far, to help piece together Bronze Age wall paintings from the site of Akrotiri on the volcanic island of Thera (modern-day Santorini, Greece) and Roman frescoes from a luxurious mansion in the Roman city of Municipium Tungrorum (modern-day Tongeren, Belgium). The computer scientists developing the system work closely with archaeologists and conservators, and the project team incorporates partners from University College London, Princeton University, the Katholieke Universiteit Leuven, the University of Athens, the University of Ioannina and the Vlaams Instuut voor het Onroerend Ergoed.

Figure 1: Modified off-the-shelf 3D scanner for rapid fresco fragment acquisition

The first step in aiding the sorting of fresco fragments is digitization. Each fragment is placed face-down on a turntable, scanned via a 3D scanning system (see Figure above), turned face-up and scanned again. Alignment is performed automatically, with the operator verifying the results and correcting any errors. A flatbed scanner is used to capture high resolution colour images and texture information of the front surface of each fragment, with the back surface being scanned once for documentation purposes. This yields a throughput of approximately ten fragments per hour, made possible through the bespoke, end-to-end design of the acquisition pipeline.

The digital surrogates of the fresco pieces are utilized in the Griphos software, which functions as a virtual table top, on which the user places fragments and arranges them for assembly, documentation or further inspection. Griphos maintains a high level of realism in an interactive viewer, which combines full 3D models and layered 2D representations. The user can have many different virtual table tops, each containing an arbitrary number of fragments, to allow multiple, competing reconstructions alongside the current physical arrangement (see Figure below).

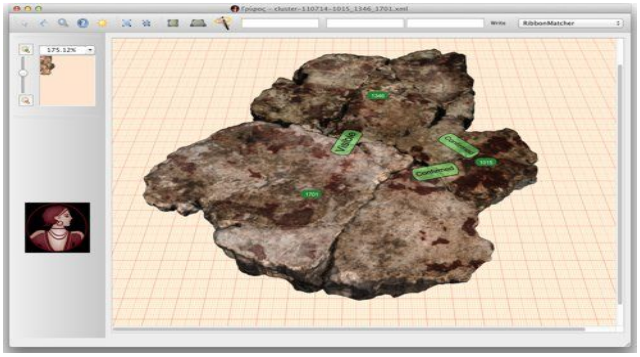


Figure 2: The Griphos application offers tools which evaluate match hypothesis amongst digitized fragments.

Griphos provides a variety of tools to study the relationship between fragments. Related pieces may be grouped together manually or snapped to each other automatically, based on computationally matched proposals, which can then be evaluated and interactively explored. An initial placement may be refined using algorithms, which provide a convenient visualization of the alignment quality, showing a colour-coded plot of where fragments touch, intersect or contain gaps. A grid is provided to aid in reconstructing hypothetical layouts and in generating overviews of fragment combinations, which saves time, as these were previously prepared by hand.

There are currently 1516 fresco fragments from Tongeren, which, statistically, means there are millions of possible match suggestions. Griphos proposed 302 matches, of which 170 were previously known, helping establish the most likely combinations for further examination. Confirming them took an expert four weeks of digital sifting, before verifying the matches against the original material. Based on discussions with expert archaeologists, we can state that the combination of acquisition- and match-browsing provides a dramatic increase in efficiency and lowers the cost of matching, compared to traditional methods.

Case Study: Developing Image capture and processing techniques to enhance the legibility of incised text

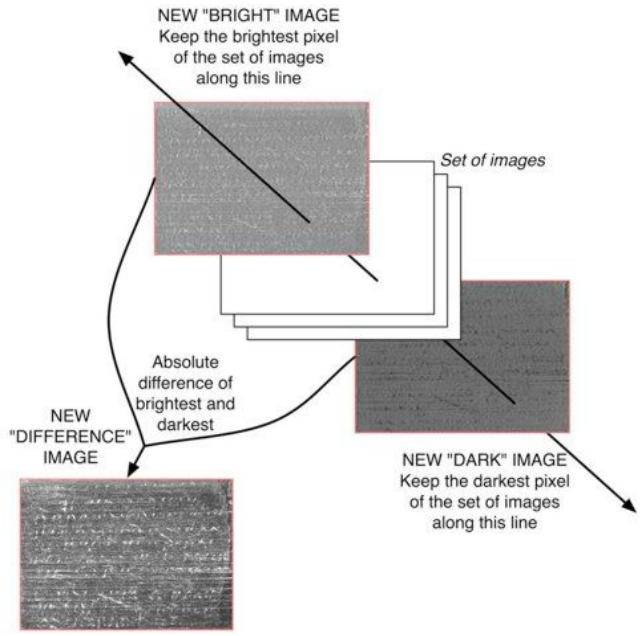


Figure 3: Example of a simple image manipulation technique on a set of images of a wooden Roman stylus tablet, where the textual information is made of the incisions in the wood.

Incised texts are often very difficult to read. Deciphering such texts draws on a wide range of expertise that papyrologists develop throughout their career. These skills are as much related to visual perception of the artefact and its text as they are to palaeographical, linguistic and historical knowledge. Computational tools can be used to aid papyrologists in their complex task, by providing alternative visual clues and evidence regarding textual incisions on documentary material. This case study presents methodologies that have been adopted to enhance the legibility of the Roman wooden stylus tablets from Vindolanda, on Hadrian's Wall, which consists of incisions left in wood through a now-perished coat of wax. The first step in deciphering these three-dimensional texts is by using advanced digitization techniques, called Shadow Stereo, and reflectance transformation imaging to capture and encode multiple images of the text under varying illumination conditions for further processing and visualization. In addition, image processing algorithms were developed to isolate the features of the text and aid in their interpretation.

Digitization of such artefacts is the first step in the development of an interpretation of the document. Observing the experts trying to interpret the stylus texts informed us as to how the tablets could be digitized, whilst retaining, in particular, one major piece of information that

papyrologists exploit when deciphering it – namely, the volumetric nature of the text. Experts who have access to the actual tablet they intend to transcribe have developed a very specific and intuitive strategy to enhance the visibility of the incisions that the stylus left in the wood. They lay the tablet flat on their hand, lift it up at eye level against a light source and apply pitch-and-yaw motions to the artefact. This effectively enhances the visibility of the incisions, by accentuating the highlights and shadows that the raking light generates; the lower the light, the longer the shadows projected by the text in (inverted) relief, created by the incisions carved on the surface of the tablet. The principle that is put into application here is the 'shadow-stereo' principle, by which concave shapes are revealed from shading, and the motion of the shadows exposes the location of the incisions (Brady et al., 2005). This process can be digitally imitated: a digital camera is affixed above a tablet, and a high resolution digital picture is taken for each one of a set of pre-established positions of a light source around the tablet. Each light position is described by an elevation angle and an azimuth angle, where the azimuth angle corresponds to an angle deviation from the horizontal in the plane of the tablet and the elevation angle describes the height of the light, with respect to the plane of the tablet. Both angles are measured from the centre of the tablet.

The collected data for each tablet was further used to digitally recreate the shadows and their motion. By adopting an appropriate model of image formation, not only can one store the set of images efficiently, without having to store each image, one can also interpolate between light positions, thus simulating lighting conditions, where a picture was not actually captured. The image model that is adopted relies on this technique of 'reflectance transformation' (Malzbender, Gelb and Wolters, 2001; Goskar and Earl, 2010). This digital model can be remotely used by experts who do not have access to the original text, allowing them to digitally pitch-and-yaw this avatar of the artefact to aid them in interpreting the text: by exploiting the play of light with the 3D nature of the textual representation, we are able to enhance the text's legibility.

The challenge of image processing and capture for ancient incised documents is multiple and specific to both the imaging method and the artefact. Our research suggests that by observing and understanding the nature of the classicist's visual expertise, we will be able to integrate prior knowledge into a model of visual perception adapted to the classicists needs, hence, supporting them in building meaning out of a pure signal – in building an interpretation of an artefact. Digital imaging and image processing can create novel, alternative visual clues as to writing contained within incised texts, assisting papyrologists in reading damaged, ancient texts.

Digitization Around Us

Digitization in its simplest form is converting and/or representing something non-digital (examples include signals, health records, location data, identity cards, etc.) into a digital format which then can be used by a computing system for numerous possible reasons.

Digitizing doesn't mean replacing the original document, image, sound, etc. Sometimes it gets destroyed (after having digitized a paper document you can destroy it or keep it, depending on, for instance, legal requirements), sometimes it disappears anyway (if we capture the sound and images in the form of video of your presentation at an event, the digital format continues to exist while your voice and physical presentation during that presentation are gone forever) and sometimes it is transformed but that's not that much about digitization in the strictest sense (if you take a picture of a building you have a digitally born representation of the building but the building is not digitized or you might have an analog picture which you scan, so it is digitized).

Digital Archives

In the context of 'physical information carriers', such as paper documents or analog, printed images, we mainly digitize by using document scanners in business (you can also, for instance, scan or simply take a picture with your mobile).

These document scanners create a digital representation (document imaging) of a scanned document, a photograph, etc. But it doesn't stop with document imaging in most cases. After all, why scan a document if you don't use the data it contains (except for archiving or, as many companies still do, for real capture after the scan)? Normally and in most business cases it's far more important that the data which capture software can retrieve from the scanned image, by using all sorts of intelligent and less intelligent capture technologies, are extracted in a digital form and leveraged to feed a workflow, a business process, a system, whatever is needed to achieve an outcome.

Online publishing

Various books, projects, scholarship and teaching material is being published online. Everybody can access these in majority cases, since it's usually not done for earning profits but for public welfare. Some examples of these are Stanford Encylopedia of Philosophy, MLA Commons offering an open peer-review site for their ongoing curated collection of teaching artifacts in Digital Pedagogy in the Humanities: Concepts, Models, and Experiments (2016) and The Debates in the Digital Humanities platform contains volumes of the open-access book of the same title (2012 and 2016 editions) and allows readers to interact with material by marking sentences as interesting or adding terms to a crowdsourced index.

Digital Photography

Photographs have always been used to convey messages, highlight products, document special events and capture images of loved ones. Digital photography has made photography affordable and accessible to everyone. Most cell phones have quality cameras and editing software built in. As a result, anyone can now take quality pictures, touch them up and share them online.

The world's first digital camera was a big box that had cassette attached to it. It was built in 1975 by Steven Sasson, an engineer who worked for Eastman Kodak. Today, about 2.5 billion people have digital cameras. Even though the standalone digital camera has changed the way we take pictures, the invention of the camera phone was even more revolutionary. In the past years, almost everything that happened on earth, from uprisings during the Arab spring to the royal wedding of William and Kate, has been captured on a camera phone.

We conducted a survey to help get an idea of how digital photography has changed the lives of people around us.

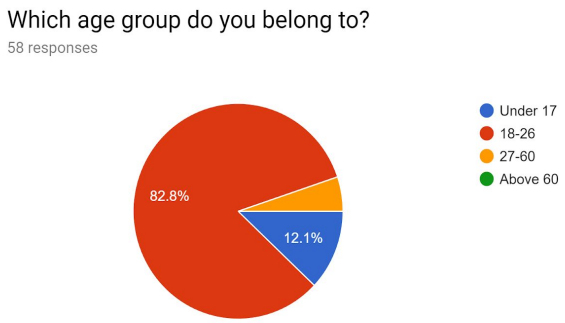
Survey

Before this age of digitization, history was woven into long and rich narratives, which were recounted with great care and thus served to perpetuate history itself. How does all of this interaction change in an era where it seems everything is recorded and stored on computers? We take thousands of pictures, but can we find the one in seconds that strikes the emotion of the special moment like first birthday photo we saved from years gone by? How does our world change as more things become digital? In the survey we conducted, we wanted to know the thoughts on these topics through some well thought questions.

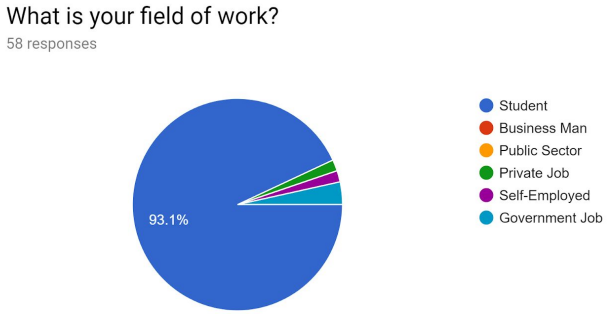
We collected data on field of work, residential area, emotional attachment, devices used for photography etc. Our focus was on determining how people from various age groups are adapting to the changing world of photography. Whether people are still interested in developing photographs or not? Do they still feel as attached to each photograph as people were a few years back? How relevant is professional photography to people around us?

Survey Results

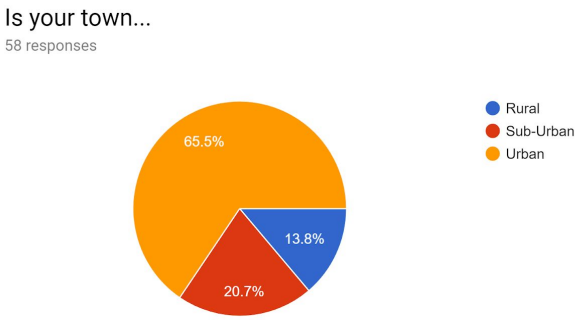
We were able to collect responses from individuals belonging to various walks of life. Following are the pie charts representing the data we collected altogether.



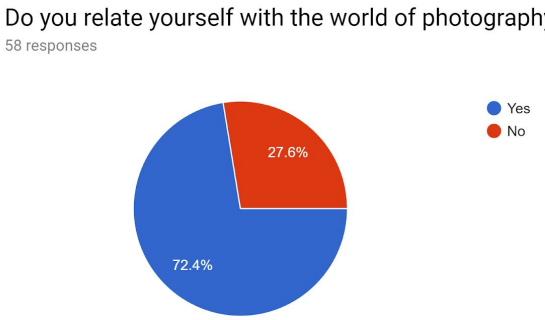
(a)



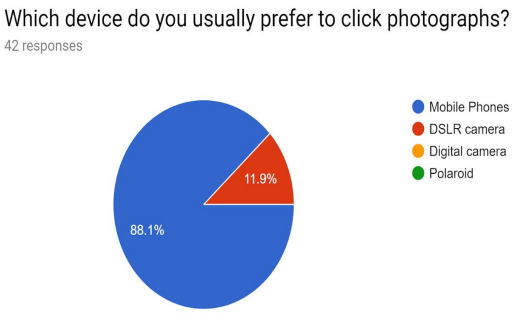
(b)



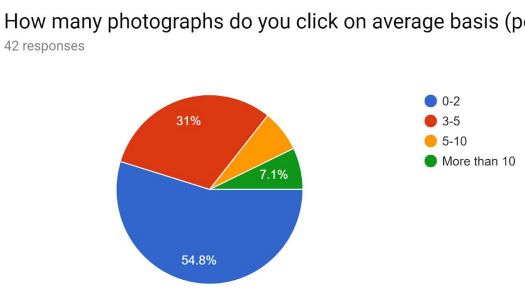
(c)



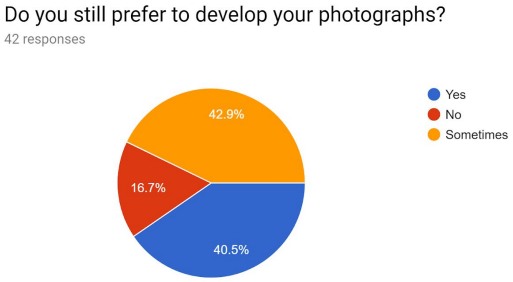
(d)



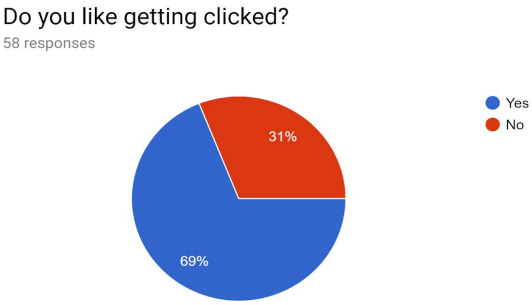
(e)



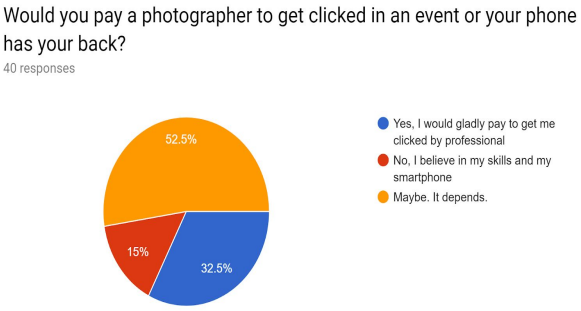
(f)



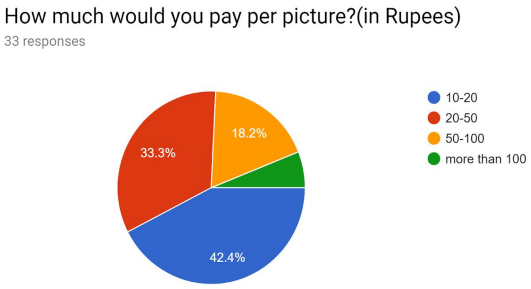
(g)



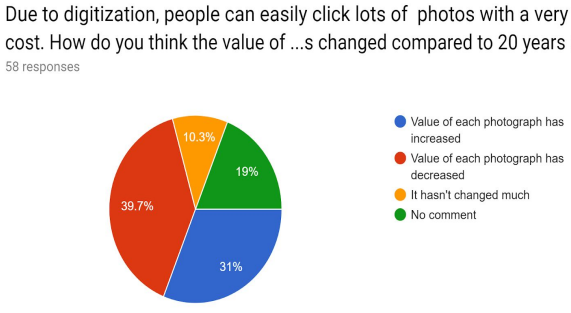
(h)



(i)



(j)



(k)

Figure 4: Pie charts representing the multidimensional data collected via the survey

Figure 4(a) is a pie chart for the age dimension. A quick glance at the figure shows that our major data set belongs to young adults. Figure 4(b) shows the pie chart for the profession as dimension. Again the majority of the data set belongs to student category. Figure 4(c) depicts the locality dimension via pie chart. The figure clear shows that our primary data set comes from urban region. Pie chart in Figure 4(d) shows how much person see themselves related to photography and it is found that every 3 out of 4 people associates themselves with photography. Figure 4(e) shows a pie chart which states that more than 70% people uses their smartphone to click pictures. Figure 4(f) pie chart shows, number of photographs clicked by a person daily and the majority of volunteers choose only 0-2 photographs. Figure 4(g) has a pie chart of whether the volunteers want their photographs developed. We collected mixed responses for this question. Figure 4(h) is a simple pie chart asking whether the volunteer likes being getting clicked and the majority said yes. In pie chart of Figure 4(i), we asked volunteers if they would like being clicked by a professional photographer. We got an ambiguous response for this i.e. more than half of the volunteers think they might want to, but that depends upon the situation. The next question we asked was how much will they the professional for the photograph and the pie chart in Figure 4(j) finds it to be around 20 rupees. The last pie chart depicted in Figure 4(k) represents the data collected for the question on the emotional attachment of each photograph they have clicked. Although more people said, it has decreased, but many of the people even voted for the increase value. Since there was not much difference between percentages, we could not conclude a single answer for this question either.

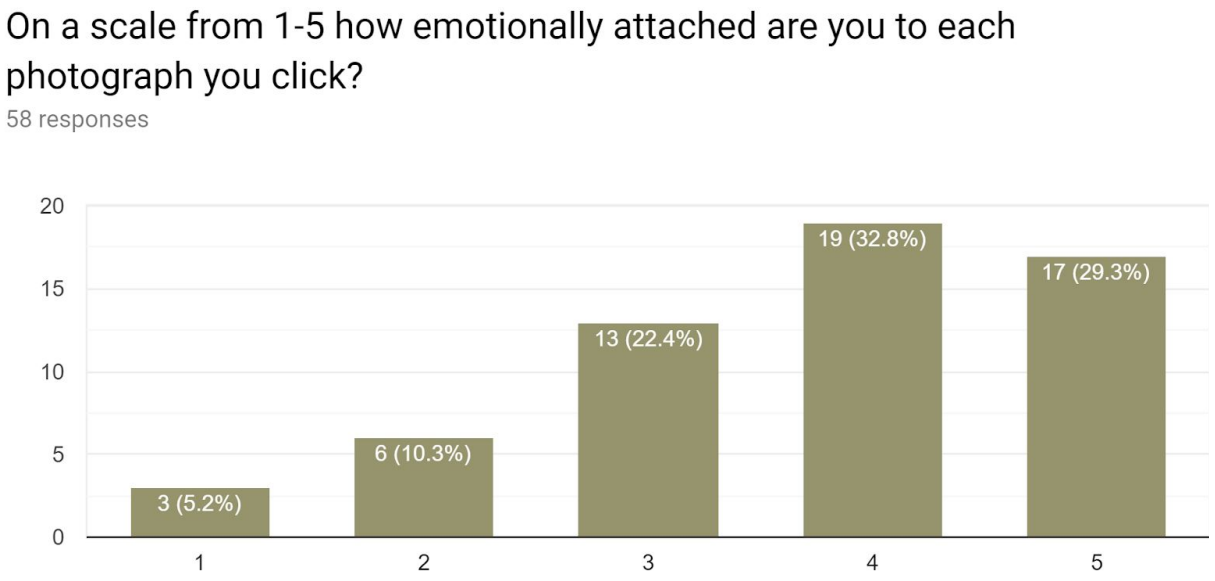


Figure 5: Bar graph representing the data collected for emotional attachment scale for a photograph

Hence we asked the volunteers to scale their emotional attachment towards the photographs they have on a scale of 1-5 (Not at all attached - very much attached) and the results we get are shown in Figure 5. It clearly shows the majority of people still consider themselves as emotionally attached to their photographs.

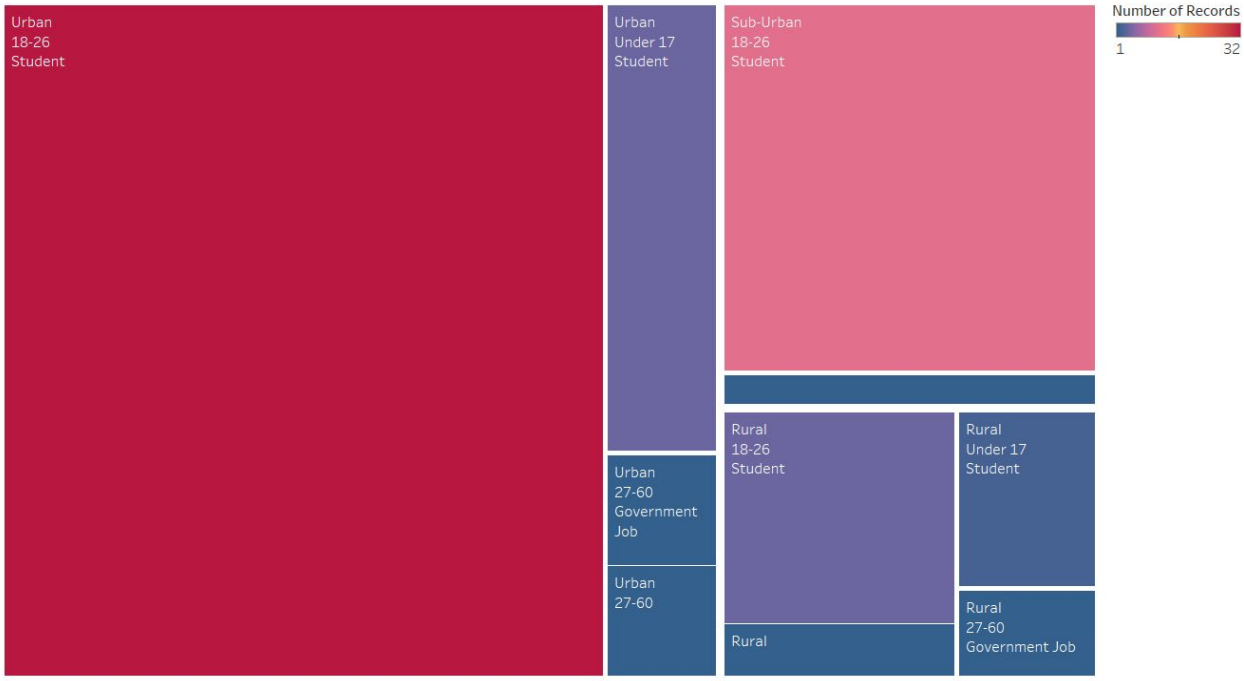


Figure 6: Multidimensional Tree map having 3 different parameters namely locality, age and profession

Further to analyse the bias in our data set we plotted multidimensional graph to check what were the dominant sectors in each dimension of our data set. Figure 6 shows a multidimensional tree map which depicts 3 different dimensions namely age, locality and profession on a 2D plane via color, size and position. Elaboration of the figure can be done to obtain that major biases in the dimensions of age, locality and profession are academic scholars or undergraduate students belonging to urban background.

Conclusion

Those who think that the ‘digital humanities’ are only about searching and making searchable large digital databases, overlook something essential. Digital humanities ask for a radical engagement with this new materiality as well as the preparedness to experiment with it. It is precisely those experiments with larger research questions about cultural complexity and cultural change, applied to larger databases, that hopefully lead to insight into, and critical reflection on, the sources we have been using in daily life. The condition technologique of our present time provides us, as a researcher, with access to much more diverse source material, allowing us to expand, focus, and broaden our research question.

That does not mean that we must ‘surrender’ to a new methodological paradigm and, with that, leave all the old behind us. On the contrary, more than ever we are challenged to confront those new sources and methods with critical interpretation and qualitative analysis. Not just that: by experimenting with digital methods, by getting to know, and work with, digital sources and by delving into the ‘secrets’ of algorithmic and computational thought, we can

better understand which dilemmas are being raised by the digitized society. By experimenting with digital methods, humanities scholars learn more about the role of Big Data in our (future) society or about the necessity of public accessibility of data. We would like to conclude with the thesis that the humanities cannot afford not to engage with ‘digitality’. Or to put it even more firmly: society direly needs the expertise of humanities scholars – their critical insights, analytical acuity, and knowledge of ambiguity and diversity – to make sense of a digital culture that permeates and directs our daily life. As academic guardians of the arts, culture, language, heritage, and the traditions of humanities thinking, we will have to engage in multifarious ways with the interrelatedness of digital technology in all kinds of cultural practices.

References

1. Brady, M., Pan, X., Schenck, V., Terras, M., Robertson, P. and Molton, N. (2005) Shadow Stereo, Image Filtering and Constraint Propagation. In Bowman, A. K. and Brady, M. (eds), Images and Artefacts of the Ancient World Vol. 4, British Academy Occasional Paper, Oxford University Press/British Academy.
2. Funkhouser, T., Shin, H., Toler-Franklin, C., Castaneda, A. G., Brown, B., Dobkin, D., Rusinkiewicz, S. and Weyrich, T. (2011) Learning How to Match Fresco Fragments, Eurographics 2011 Special Area Track on Cultural Heritage, April , http://doi.acm.org/10.1145/2037820.2037824.
3. García Castañeda, A and Brown, B and Rusinkiewicz, S and Funkhouser, T and Weyrich, T (2011) Global consistency in the automatic assembly of fragmented artefacts. In: Dellepiane, M and Niccolucci, F and Pena Serna, S and Rushmeier, H and Van Gool, L, (eds.) 12th international symposium on virtual reality, archaeology, and cultural heritage. (pp. 73 – 80). Eurographics Association: Goslar, Germany.
4. Goskar, T. A. and Earl, G. P. (2010) Polynomial Texture Mapping for Archaeologists, British Archaeology, 111 (111), 28–31.
5. Malzbender, T., Gelb, D. and Wolters, H. (2001) Polynomial Texture Maps. In Pocock, L. (ed.) ACM SIGGRAP'01: proceedings of the 28th annual conference on computer graphics and interactive techniques held on 12-17 August 2001, Los Angeles, CA, USA. organized by ACM.
6. Shin, H., Doumas, C., Funkhouser, T., Rusinkiewicz, S., Steglitz, K., Vlachopoulos, A. and Weyrich, T. (2011, forthcoming) Analyzing and Simulating Fracture Patterns of Theran Wall Paintings, ACM Journal of Computing and Cultural Heritage (JOCCH).
7. Tarte, S., Bowman, A., Brady, M. and Terras, M. (2010) Image Capture and Processing for Enhancing the Legibility of Incised Texts. In Holappa, M. (ed.O , Eikonopoiia: Digital imaging of ancient textual heritage: technological challenges and solutions held on 28–29 October 2010, held on 28–29 October 2010, organized by the Centre of Excellence (Academy of Finland)Helsinki , www.eikonopoiia.org/files/Eikonopoiia-2010-Proceedings.pdf .
8. Toler-Franklin, C., Brown, B., Weyrich, T., Funkhouser, T. and Rusinkiewicz, S. (2010) Multi-Feature Matching of Fresco Fragments, ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2010), 29 (5), 185: 1–185: 11 , http://dx.doi.org/10.1145/1882261.1866207.
9. https://master.epfl.ch/programs/digitalhumanities/
10. https://cis-india.org/raw/digital-humanities-in-india