Fuchsgruber, Lukas & Bianca Ludewig: 16 Questions to Mariana Lanari of Archival Consciousness. In: Reclaim your Archive! Online article series on reclaim.hypotheses.org, published by Art Doc Archive Berlin, March 2023.

Author

Mariana Lanari is an artist and PhD candidate at the Amsterdam School of Heritage, Memory, and Material Culture (AHM) at the University of Amsterdam. In 2019 she cofounded Archival Consciousness in collaboration with graphic designer Remco van Bladel to investigate new publishing methods for cultural libraries and archives. Lanari is a member of the advisory board of De Appel Mobile Archive (NL), and of DAAP – Digital Archive of Artist Publishing (UK) and Printroom (NL). She is a member of ARIAS, Amsterdam Research Institute of the Arts and Sciences, and of TARR - Third Cycle Artistic Research of the Rietveld Academy. Her work has been shown at the Stedelijk Museum Amsterdam, Van Zijl Langhout Gallery, FLAM - Festival of Live Arts Amsterdam, Arti Amititae, W139, Van Abbemuseum, Transmission Gallery (Glasgow), Casa do Povo (São Paulo), de Appel Amsterdam, among others. Archival Consciousness is supported by Creative Industries fund NL, Digital Culture program and MediaFutures, received funding from the European Union's framework Horizon 2020 for research and innovation and is part of ACKnowledge, Art Routes from the Dutch Research Agenda 2020 (NWA/NWO).

art doc archive

Art Doc Archive is developing a prototype for the centralized archiving of the materials self-documented by Berlin-based artists on websites and in social media. The "Reclaim your Archive!" series of articles accompanies the project and is edited by Bianca Ludewig. All texts can be found on the project's blog: reclaim.hypotheses.org

Art Doc Archive was funded by:

Senatsverwaltung für Kultur und Europa | BERLIN

16 Questions to Mariana Lanari of Archival Consciousness

Interviewed by Lukas Fuchsgruber & Bianca Ludewig of Art Doc Archive

1 Lukas Fuchgruber: Can you tell us a bit about your work background and your project Archival Consciousness, what is your experience with digital archives?

I am an artist and researcher with a background in Political Sciences. I'm doing a practice-based PhD at the Amsterdam School of Heritage, Memory, and Material Culture (AHM) at the University of Amsterdam. Since the early 2000s, I have been involved in the design and development of databases used in the production of exhibitions to manage loans, transportation, info for subtitles/wall texts, and info about artists and organisations, such as the biennial of São Paulo (2012). I also worked as a book editor in a publishing house where I had contact with another type of management system used to control the entire book lifecycle, from the number of pages to print run budget, price composition, storage costs and sales. As an editor, I often went in person to consult documents in physical archives. And I'm part of a generation that grew up using paper dictionaries.
My work in archives and libraries in cultural organisations results from these experiences and aims towards a critical intersectional feminist decolonial standpoint in relation to information management systems. Since 2015 I've been researching forms of display, access and new ways of reading in libraries and archives in cultural institutes considering the library as a laboratory to investigate the use of computational methods aiming at digital literacy (starting by my own), to develop a more informed critical view towards digital technologies.

In 2019, Remco van Bladel and I founded Archival Consciousness to work in collaboration with archives and libraries to develop web applications approaching the archive with an editorial mindset. We use concepts and techniques from bookmaking to investigate ways of publishing from a database perspective to make these sources of information more readable, resilient, and accessible to their communities. Remco co-founded publishing house Onomatopee in 2006. For the past 15 years, he has been involved in making books, websites, and platforms such as WdW Review, e-flux SUPERCOMMUNITY, and the 13th Gwangju Biennale. He has witnessed the impermanence of websites that are made for events such as biennials, bringing together valuable content that, in a few years, become inaccessible, and the problem of information loss due to website migration.

2 Lukas Fuchsgruber: What do you think are the main challenges when you are adding more linked information to existing datasets? Which methods worked well and why?

We consider exporting datasets as a form of publishing the database, in curated or total exports in a human and machine-readable way. For datasets to be useful, we add context and describe it in detail, how, when by whom the data was collected, what it can be used for, etc, following the recommendations described in the paper *Datasheets for Datasets*[i] adapted to our use case. There are many challenges involved in aggregating data to existing databases, especially if it includes the participation of the community.

The first challenge is to create a system to make this an ongoing process. It's possible to work with static archives, but we mostly work with living, dynamic archives, where the task is archiving in the present as well as maintaining the archive over time. Another challenge is to create a method and break it down so that more people, specialists in their own domains, can contribute to the process of making data without necessarily having technical skills.
Regarding methods, in the case of Biblio-graph, we started data aggregation during the exhibition *Catching Up In The Archive*, in spring 2022, showcasing the entire archive of de Appel. Around 3.5 thousand people visited the show and could add data to Biblio-graph. This was an opportunity to test and present the system and have intense participation and data aggregation. It created a habit in the team of the archive, and after the exhibition, the system continues active in the archive.

Biblio-graph functions like a copy machine in the archive, where visitors make copies of the pages they want to read later. The pages are published on biblio-graph.org and add to an overview of the archive. Data aggregation is not done as an end but in parallel to the use of the archive. To capture images of objects, we made a Mobile Archive Unit. It's a table with lights, an RFID reader, and a phone with a custom application. Every item in the archive received an RFID tag with its corresponding id in the database. When a person puts a publication on the table, the app on the phone recognises the id of the book, and the reader can capture images of pages that will be stored in the database in the record of the publication. And it is published on biblio-graph.org in real time. It is also a way of broadcasting live from the archive. This method creates a shortcut in the process of digitisation, renaming of files and manual uploads to the database. Biblio-graph currently has a timeline view, with a history of the books being read in the archive. There is a map with an overview of the pages. And each page is an entry point to a graph view that shows the network of people, organisations and publications. We don't intend to digitise entire publications or to make a reader, the application is intended for data enrichment, since most of the publications are not available online, research, reuse, and visualisation. There are other ways to find entire books, such as in the physical archive in this case.
Another process of data aggregation is annotation/labelling. We want to harvest data manually and gradually, as a form of curation made by the community, the readers, researchers, curators, and staff of the archive over time. We are experimenting with NLP (Natural Language Processing) tools and methods that are used for labelling datasets to train machine learning models, but with a radically different use case, mindset, and values. These tools allow tech companies to hire thousands of people that are exploited and underpaid to annotate datasets simultaneously, often without knowing what the data is and what it will be used for.

In our case, we are not building models, we are not interested in any type of prediction, we are not making commercial applications, and we are not hiring people to do the work. The annotation is done by staff, researchers and the community of the archive for their own benefit, of being able to process and read large amounts of content. We are using basic features of data annotation tools to create a corpus based on the collection of the archive while reading or researching. Only when a word is manually highlighted then we use Natural Language Processing methods to identify entity relations (NER), dependency and coreference resolution, and mapping other occurrences of the same terms, storing these terms to pre-annotate texts that are added later.
If a system of data aggregation is part of an existing archive, another thing to keep in mind is that such a system could lower the workload of archivists in certain tasks, like metadata input and enrichment. On the other hand, it could change and perhaps increase the work of the archivists, to manage, verify, and curate, fine-tuning this balance is another challenge. Many institutes don't have archivists, in these cases, one or more people are responsible for the tasks of archiving,

usually on top of their main activities. We find it important to acknowledge the labour and routine involved in archiving. Doesn't matter how much you automate, an archive will still demand archivists, maintainers, and caretakers. Therefore, it's important to analyse what the effects are that automation on any scale might bring and how it affects people.

»My personal interest in this field is to work with the "messy" data, to prepare and sculpt datasets. «

3 Lukas Fuchgruber: We face the challenge of messy data coming from a variety of sources (websites of artists, social media feeds of artists). In your experience what are the best approaches when faced with diverse or unstructured data?

My personal interest in this field is to work with the "messy" data, to prepare and sculpt datasets. There is no shortcut, the process is time-consuming, delicate, and involves manual labour. It is like editing, it requires patience, an eye for details, translation, interpretation, and decisions in every step. But it pays off, because the bulk part is done only once, then it is maintenance work. I'm not in favour of the term cleaning, nor of excessive cleaning and normalising. On the contrary, there is a lot to learn from the nuances in the data. I prefer a soft approach to cleaning and correction, and to call it data preparation, as a process of understanding the data and widely documenting the process so that other people can understand the choices that resulted in a dataset. Messiness is somehow misleading because it suggests that a database should be tidy, which cannot be the case in systems that are accumulating data created manually for decades. The messiness is the human factor, and thanks to all the people who patiently input data onto databases over the years, we now have amazing and extremely valuable datasets.

Data preparation can be approached in stages by mapping all data sources to one data model that expresses the network of relations within that collection, and then proceeding field by field, adjusting and preparing the data. It is a rich process that gives many insights into the data and into the archive. I see it as a form of reading and interpreting the archive. The process informs how to display this information and opens a lot of possibilities in terms of ways of publishing as well. Different types of visualisations, such as graph visualisation, help in spotting issues in data, that's why it is nice to make it as an iterative process and not just a stage at the start of the project. I like it, it is like editing a book in volume and detail, but in a non-linear way and using different tools. That is why I find data/digital literacy important so that more people can take care of their data and participate in works that involve data which are in great demand nowadays.

This aspect of data collection, curation, and preparation is precisely where we can influence and contribute to digital applications that rely on data, from the perspective of the humanities. It is where the broad fields of the social sciences and engineering meet. Very little attention is given to how and by whom datasets used in Machine Learning models are created. These systems are conveniently portrayed as black boxes when in fact, whatever comes out of it is based on what came in in the first place. The black box is the computational part, the data is not a black box, and people are conveniently not held accountable for the collection and preparation of data, when these systems produce harm, then it is a black box.

4 Lukas Fuchgruber: In our project we are parsing websites and social media feeds of artists for information about their works and practice. We mirror their websites and feeds and add our analysis and visualization. One question related to the project management

and prototyping, especially the early stages: From the experience in your project, what do you think is the best approach to structuring such exploration and extraction? How do you conceptualize connections in diverse material, what made the most sense in your project? How did you approach the conceptualization, how did you set the focus and decide on the tools you want to use?

In our project, we start by making an ontological model of the institution that is behind the archive, in conversation with the team of the cultural institution, to understand and map the context and particularities of the institution. Ontology is a conceptual model, a form of knowledge representation within a specific domain. It is an explicit description of classes (concepts), their properties, attributes, and relationships, in a machine-readable way, that are enough to capture the essence of an institution.[ii] The upper-level ontology, which is a more generic layer, helps to load several databases in the same environment, and the particularities of each organisation help to give more context to the data.
Then we look at the available data and make a data model that will populates classes of the ontology. The data model is a simplified version of the ontology if you superimpose both models. We look at the current website, legacy websites, and databases in relation to the data model, in terms of how to map the fields from web pages to the data model. We structure the content as much as possible in the ingestion. Blocks of texts like event descriptions or pages will be structured at a later stage by researchers and staff, using data annotation tools. It is a process of translation, and it is a delicate operation because it changes the context of the content. It is important to widely document the process and decisions.

5 Lukas Fuchsgruber: We see large potential in community-based research, where the question and perspectives are contributed by people who have their own practices of archiving and documenting. From your perspective, how can this be sustainable? How to care for the network/community around archival projects? What role plays community in Archival Consciousness and what are the lessons learned related to that?

We think sustainability needs to be a collective effort with several people sharing the responsibility. In our case, the only way to make it sustainable is to share the basic infrastructure [database, CMS, frontend] between several organisations. That is an advantage of working with cultural institutions such as Framer Framed that is already working with several communities. Our model is that every institution implementing Biblio-graph can transmit the method and offer the infrastructure to host other institutions or archives in their networks.
In our experience, we see that different archives complement and strengthen each other. We have a lot to learn from archivists regarding methods and routines, and the type of technology we use in the backend should become relatively easy to share. In our project, the technology is open source, but the method needs to be transferred from person to person across institutes. So, the community has different roles in the making of the digital archive, through data aggregation, and then in sharing the method and infrastructure with other communities as a strategy of preservation and maintenance. But it is too early to say if this model will work. So far, we are developing one step at a time, one grant at a time, and one institution at a time. And this is covering the most expensive part of processing the data, prototyping, development, and implementation. Once this is done, if multiple archives become part of it, maintenance costs will be reasonable.
At the same time, considering the environmental costs of digital applications in general, we do not think of the system being sustainable in the long run. We see this system as transitory and relevant

for the preservation of the knowledge in the archive in the form of comprehensible datasets and their explanations, ontologies and models that can be published and even printed like a book. That is why we place much attention on ways of exporting data and composing datasets so that they become portable and available in different formats.

»We don't think of the system's sustainability in the long run«

6 Lukas Fuchsgruber: How do you see your project Archival Consciousness in relation to platform politics? For example, the problem of archival communities existing on corporate platforms like Facebook, Instagram and Google Docs, and their interactions being fed into the business modes of surveillance capitalism. How does your project challenge this influential mode of digital communities?

I cannot say much about this because platform politics is not my field of study. We want to make our own tools outside of commercial platforms, building our models and algorithms from scratch, to make a system with less automation and that is not based on prediction, using existing technologies such as graph databases, ontologies, and NLP, without commercial purpose, and without tracking people's data. Except for graph databases that became more available lately, these technologies are not new. We try to use these technologies specifically for learning and making large quantities of information readable and useful for communities, and at the same time to make demystify the technology.
I see this as a privilege though. There are few examples (and some examples below) of communities building their own infrastructures, but this is not possible for many communities outside the Global North that unfortunately must rely on the corporate platforms to organise and archive.

7 Bianca Ludewig: You pose in a recent research paper, "How to Read a Library", this following question ... what is the answer? – "Can machine learning and AI be used in a library to enhance reading and promote access instead of being used for targeting advertisement and surveillance? "

From a technical perspective it is certainly possible, and there are amazing examples of communities making and using machine learning, such as the Māori language revitalisation project by the Hiku Media[iii] or Masakhane, a grassroots organisation whose mission is to strengthen NLP[iv] research in African languages, for Africans, by Africans.[v] It's extremely problematic that Big Tech like Meta and OpenAI hijack these projects, scraping their content to train their models and sell it back to us.[vi] It's challenging in many ways, how Big Tech finances and offers education programs to train developers according to their ideologies, which reinforces ways of doing and types of use cases, making it challenging to explore the technology using methods that are not commercial and not intended to optimise profits.
As Dr. Abeba Birhane very well articulate in her influential AI ethics paper "Algorithmic Injustice: A Relational Ethics Approach,"[vii] machine learning models are behind systems with huge social impact that rely highly on data, the backbone of these systems is data. In fact, the technology used by these systems has been the same for decades. What changed is the availability of substantial amounts of data. Since libraries and archives are information spaces, they can be instrumental in education, learning and understanding data and systems that rely on data. That is why cultural archives that are conscious of their value for their communities and the next generations might

want to go beyond taking responsibility and agency towards their data to also promote digital literacy and a critical view towards these existing systems.
Archives and libraries in cultural institutes are spaces of resistance and dissident voices. They have rich and important collections that are not available digitally. Many of these archives are facing a precarious situation, and many are closing. Also, an entire generation of archivists are going to retire soon, and archiving is not solved by technology or by digitisation alone, it needs people, methods and routines.

8 Bianca Ludewig: You also state there that working with data & databases of archives: "involves working with the database managers and IT professionals that are often not part of the team, but external companies". What does that mean if we think of the preservation and maintenance of digital archives?

I cannot generalise since it varies case by case. When an institute wants to open or share their data. Sometimes, the person that decides about sharing data is not the same person that created and maintains or that has access to the database. Sometimes the archivist works with the database but is not responsible for managing it and cannot give access to it. The database might be managed by an IT professional outside of the organisation. In our case, getting access to data in an organisation is always part of a long-term collaboration with several people and involves trust.
In terms of the preservation and maintenance of digital archives, it depends as well. We cannot talk about the digital archive in general because it can mean different things. We think in terms of living archives, archives that are dynamic and active, part of a community that produces publications or events, like de Appel, Casa do Povo, and Framer Framed, to give some examples. In these cases, preservation and maintenance are part of the activities of the organisation, which is also busy archiving what is happening in the present. The archive of past events and publications is also alive, records are being enriched or amended and are activated through relations with other records. We don't work that much with static archives. But even a static archive, if accessible for consultation and research, needs maintenance.

9 Bianca Ludewig: I like especially the idea that digital data created by archive users becomes part of the commons and they thus become co-owners/ creators. Is this imaginable for all digital data created in general? What is your understanding of public and commons?

I don't think it is imaginable for all digital data, and perhaps it is not desirable. I don't like to address data in general, I prefer to consider case by case, and in context. In our system, data is added and will be annotated by the community around specific archives in cultural institution. We acknowledge that when people contribute to the archive, it is a form of labour, so in exchange, the archive doesn't become the owner of the resulting data, and it becomes public. We at Archival Consciousness also don't own the data. If we aggregate data through the community, then the community is the owner of the resulting data, meaning that it cannot be commercialised in any way. To avoid this, the archive needs to agree that if they want to have community participation in their system, they need to agree not have ownership over the results of data aggregation.

I do not have a solid understanding of the public and commons. We see in our practice that not all data is suitable to be made public. There must be distinct levels of accessibility and it is up to each archive and the community to decide what is publicly available or not. One thing is ownership and authorship, and if it is publicly available, open or closed, is another topic. Because this addresses

only part of the problem of data theft, as in the case mentioned above, of big tech scraping whatever they want, packaging and selling it back without any acknowledgement or compensation for the community. This case below illustrates this, I'm quoting from the commentary of Dr Abeba Birhane about the project of masakhane community mentioned above, on Twitter, from August 3, 2022:[viii]

"over a year ago, i commented on how big tech might take African language datasets, collected and curated by underfunded and overworked African researchers and I was dismissed as 'alarmist' 1/

today i learnt of a project inside google on African NLP/AI that has emerged on the backbone of these datasets 2/

similarly, you find companies like Meta building language models (including for African languages) using datasets that are put together by communities such as masakhane. Do they fund any of these disfranchised African researchers? NO!

what's worse, companies like Meta take these 'open sourced' datasets and licence them (or the tools they are building from them).

this infuriating. Not only they control the datasets that come from marginalised communities, they also control the narrative, language and knowledge system and then sell it back to us.

This is a big problem & needs immediate regulatory attention. I know many African researchers with innovative tools/methods/datasets but are not willing to share them as they know they will end up being stolen &/or appropriated by big tech/those w resources & in positions of power.

the push for 'open sourcing' without appropriate regulatory and structural safeguard will only benefit the already powerful and in possession of resources. end/

10 Bianca Ludewig: Could you explain this statement of yours: "the logic governing both the physical and online archive is open to debate"...?

I said this about the data model and the ontology. Understanding the model is key to understanding the data. And the model, in our case, is made in conversation with the archive team and the institution and based on existing databases and websites. We agree on an ontology and a data model, and it is to a certain extent, open for debate because there is not one right way of modelling data. Ontology is helpful but not ideal in the humanities, where things cannot be defined by logic and needs to be taken with a grain of salt. In any case, there is always a model behind the database, reproducing a certain worldview.

»In a time of so much hype around AI, we would like to take the longest route of education and literacy as a form of resistance. «

11 Bianca Ludewig: You emphasise in "How to Read a Library", how human intellectual labor, "involved in manual annotation, is what animates the online machine". There seems to be so much human labor involved in digital processes which is hidden to the public. Why is it important for you to highlight this?

Many tools advertised as artificial intelligence heavily rely on people in the background. Including social media that only works thanks to thousands of moderators that are exposed to the most violent content. Data labelling is the starting point of many AI application, data is the backbone of machine learning models that are behind several social control systems much beyond social media, such as immigration, health care, insurance, retailers, and sorting out CVs, to list just a few. We need to have at least a basic understanding of how these systems work, how data is produced and processed, and what the limitations are. In this sense, we see the archive and library as a laboratory to experiment with data and digital technology, especially because the provenance of data is right there and accessible.
In a time of so much hype around AI, we would like to take the longest route of education, digital and data literacy as a form of resistance. It is crazy that people want a robot to write for them, we prefer to invest in better tools for people to read and write. Writing is a form of expression, what is being expressed when chatbots are writing on our behalf?
Whatever the robot writes are based on the content that it was trained on. We need to position critically in relation to this, and one of the ways is by understanding the role of data and of people behind these systems.

12 Bianca Ludewig: Archiving and preservation is often dependent on power and money. You understand art or knowledge preservation (archiving) as a collective endeavor, how can this process become collective and fair? Especially in the digital realm where bias is an ongoing problem. In a way that not only the knowledge produced by historically dominant groups and institutions is preserved, but in its diversity?

Indeed, archiving and preservation are instruments of power. But community archives can also be an important instrument of counter-power. If a community is conscious of the relevance of its archives, it will work collectively to maintain it, not out of choice but by necessity to keep its history and memory alive. It is precarious, and many small archives get lost or constantly risk disappearing. It is interesting to note that, as Shelley Angelie Saggar wrote in the *Decolonial Dictionary*, the mission to "preserve" from the perspective of power is sometimes at odds with communities' visions for the use and futurity of their material culture.[ix]

Biblio-graph, the web application we are developing, is built for archiving as a collective process for several reasons. One is to include actors and voices of the communities, the staff of institutes, researchers, and the public. But to promote a basic and hands-on digital and data literacy. What you see on biblio-graph.org starts in the physical archive, by digitising documents and publications, and later annotating. This system illustrates *a* process of data creation, by which a person chooses a publication, and then decides about which specific pages of that publication to capture, it is stored in a database as part of the record of that publication and published online. The process is very simple, but it becomes clear that in every step there is a choice, and that things are constantly being filtered out. Later it will be annotated, and what is annotated and what is not is a personal choice. If the community is part of the making of the archive, they will understand how that archive is made and how other archives are made. They can shape the archive, have a more active voice, and ownership.

Another problem of archives and libraries is that they operate with a small team, by allowing community participation in archiving, the team can get some extra hands. There are several organisations that do not have an archivist, then the fact that the tasks of archiving can be distributed among the users of those materials can increase the chances of preserving them. We are interested in exposing how data collection works and how datasets are made, and what types

of information and knowledge we can gain from visualising these datasets as a network and what are the limitations. In a way that any conclusions made through data analysis or visualisation can be verified because there is provenance, the sources are available in the archive. All these details are crucial for the accountability of any decision made based on data. A system that cannot disclose the sources of data used to train their model cannot be trusted.

13 Bianca Ludewig: You highlight in "Reading as Sculpture" that it is "increasingly relevant that artists, intellectuals, and designers claim responsibility in the creation of digital tools". Why is this so important and how could we get there? Does that mean that we all have to become programmers?

This is a quote from Johanna Drucker: "The design of digital tools for scholarship is an intellectual responsibility, not a technical task. Unless we [artists and] scholars are involved in designing the working environments of our digital future, we will find ourselves in a future that doesn't work, without the methods and materials essential to our undertakings. We must be committed to designing the digital systems and tools for our future work. Nothing less than the way we understand knowledge and our tasks as scholars are at stake. Software and hardware only put into effect the models structured into their design."[x]
I do not think we all need to become programmers. But we cannot afford anymore to completely ignore how digital technology works. I we must stretch the border between what we can do and where the domain of engineers starts. Data, databases, and data models, for instance, are areas in that we, in the humanities, can participate more. It requires minimal or no coding skills. We don't need to develop a database to be able to manipulate it, find problems on it, and challenge assumptions based on data. The AI hype feeds from this ignorance about how it works and what the limitations are.

14 Bianca Ludewig: Apart from exceptions art is precarious, writing is precarious, culture at large is precarious – how can knowledge or art outside of the mainstream or institutionalized settings be preserved and secured?

We are dealing exactly with this aspect of precarity. Our strategy is to display this state of precarity in archives and libraries in cultural organisations and to display the collections as well, and to bring attention to it. We are a small studio, and we are developing infrastructure for important institutes like de Appel and Framer Framed. This fact alone reflects the current state of precarity not only of the archives but of systems that can do this in a non-commercial way. We are developing this system through research grants, but infrastructure cannot be seen as a project, it needs continuation, it needs to be maintained in the long run. The only way of maintaining this is to share the maintenance among many institutes.
We started working with the digital-born archive of Framer Framed, and this year we might implement this system to 10 other organisations, including art publishers and magazines, archives in cultural institutes and digital-born archives. With more archives sharing a system, the interconnectedness becomes more visible, and collaborations between the archives that go beyond the infrastructure already started appearing spontaneously. It needs to be a group effort. In this sense, the fact that we are a small studio and not part of one of the archives is also a facilitating factor. We are building from the bottom up, one institution at a time, in an iterative way, adding new functionalities on the go.

15 Bianca Ludewig: It seems to me that one of your aims is to show how interlinked archived objects are with other objects, people, contexts. You also talk of "finding narratives through linking"... This is also something important for Art Doc Archive, how can thick contextualization in an archive be obtained?

Each archive has a story to tell. We do not start with the data, but with the history of how that collection came about, why it started, when, where, by whom. This sets a context for mapping materials in the collection, and translating the collection as data, also structuring parts of the content of publications, audio, or film, and not only metadata. The process of data collection and preparation provide insights that can be formulated as queries or questions, that can be stored and displayed.

»We are thinking about publishing as archiving«

16 Bianca Ludewig: In one text you also refer to "hybrid archiving", what does the term mean as a praxis?

I'm not sure if this is a good term. The term "hybrid" gained a different meaning after the pandemic, referring to events simultaneously online and in person. In the case of digital archiving, we are thinking about publishing as archiving, inspired by physical archives in cultural institutions and their application-agnostic databases. A database that not only serves a website but first stores data and metadata about the content of pages and makes the content retrievable. What we are referring as a digital archive is a repository to map, visualise, browse, and retrieve information about items in a collection in an accessible, and reusable way.

These two processes, publishing online and archiving, are usually done in different moments and environments. Publishing as archiving means that the system is built in a way that that the making of one (publishing) produces the other (archiving). The reason to do this is the volume of content produced and the lack of infrastructure and personnel to make and maintain an archive. Archiving as publishing is also an interesting concept, considering the bibliography genre of publications, that could be interpreted as a description and listing of items in an archive. Publishing datasets that represent a whole collection, or parts of the collection as a documentation about an archive – like a phonebook that lists every person, organisation, object, event, and how they are related – can be seen as an act of archiving in the sense that you can reconstruct an overview of the collection.

Both: Thank you for these insights!

ENDNOTES

[i] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets." *Communications of the ACM* 64, no. 12 (2021): 86-92. Accessed March 23, 2023. https://www.fatml.org/media/documents/datasheets_for_datasets.pdf

[ii] Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001). Accessed March 23, 2023. https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[iii] Finn, Aoife, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. "Developing a Part-Of-Speech tagger for te reo Māori." In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 93-98. 2022. See also "Te Hiku Media" Accessed March 23, 2023. https://tehiku.nz/

[iv] Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola et al. "Participatory research for low-resourced machine translation: A case study in african languages." *arXiv preprint arXiv:2010.02353* (2020). Accessed March 23, 2023. https://arxiv.org/abs/2010.02353

[v] Marivate, Vukosi, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. "Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi." *arXiv preprint arXiv:2003.04986* (2020). See also " Masakhane, A grassroots NLP community for Africa, by Africans" Accessed March 23, 2023. https://www.masakhane.io/

[vi] Mahelona, Keoni, Gianna Leoni, Suzanne Duncan, Miles Thompson. "OpenAI's Whisper is another case study in Colonisation" The Papa Reo, Jan 24, 2023. Accessed March 23, 2023. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/

[vii] Birhane, Abeba, and Fred Cummins. "Algorithmic injustices: Towards a relational ethics." *arXiv preprint arXiv:1912.07376* (2019).

[viii] Abeba Birhane (@Abebab). *Twitter*, August 3, 2022. https://twitter.com/Abebab/status/1554869921778974 74

[ix] Saggar, Shelley Angelie. "Decolonial Dictionary". Accessed March 23, 2023. https://decolonialdictionary.wordpress.com/

[x] Johanna Drucker. 'Blind Spots: Humanists must plan their digital future.' in The Chronicle of Higher Education April 3, 2009. https://www.chronicle.com/article/Blind-Spots/9348