The research leading to these results has received funding from the European Commission under grant agreement No. 621610-EPP-1-2020-1-NL-EPPKA2-KA.

D3.2 - Second Version of Teaching Modules

Version 1.0 (*Final*)

July 20, 2023

Co-funded by the Erasmus+ Programme of the European Union

Grant Agreement number:	621610-EPP-1-2020-1-NL-EPPKA2-KA
Project acronym:	MediaNumeric
Project title:	MediaNumeric: Alliance on Multimedia Data-Driven Journalism & Media Production
Funding Scheme:	ERASMUS+ KA2: Cooperation for Innovation and the Exchange of Good Practices - Knowledge Alliances
Project coordinator name, Title and Organisation:	Johan Oomen, The Netherlands Institute for Sound and Vision
E-mail:	joomen@beeldengeluid.nl
Project website address:	http://medianumeric.eu/

Revisions

Version	Date	Author	Reviewer(s)	Changes
0.1	30.02.2023	Ewa Korzeniowska	Kuba Piwowar, Rachel Somers Miles	Prelim Review
0.2	12.07.2023		Marie Hospital, Susannah Montgomery	Partner Review - minor feedback
1	20.07/2023	Ewa Korzeniowska		Final version - minor corrections

Table of Contents

1. Executive Summary

This deliverable describes the activities carried out during the first iteration of a Summer School of the MediaNumeric programme held in The Hague in June/July 2022, Work Package 3 (WP3) by the seven consortium partners and describes the results achieved by this work package. It outlines how the programme's structure reflects both the Needs Analysis and State of the Art reports' findings, refers to all materials used during The Hague programme and the idea behind the case study. The case study was the backbone of the entire programme: during the Summer School, the students worked on the topic of diversity in the music industry, using knowledge and tools they were provided during classes. As a result, each group created their own 'pitch' presented to the consortium representatives and fellow students, based on a topic of their choice that referred to a broader theme of the music industry and social phenomena connected with it. Each lecture and workshop was designed in a way that provided the students with skills necessary to complete all tasks, starting from data gathering and analysis, through storytelling and visualisation, and finishing with debunking disinformation exercises.

2. Introduction

This deliverable consists of a detailed description of the Summer School that was held in The Hague in June/July 2022. It entails a full syllabus for all three modules of the course in data journalism and digital verification, descriptions of all lectures and workshops, links to accompanying presentations, data sets, handouts, and additional materials that were provided to students during the six-day programme. It also includes the three quizzes that were designed to verify knowledge gained by the students throughout the course, as well as links to the final presentations of the case studies that students put together during the week. The document also outlines the study visit to The Netherlands Institute for Sound & Vision (NISV). The final section details the consortium's plan for recording the classes with video and outlines options being discussed regarding the digital format of this training course. The report concludes with detailed notes taken during a debriefing among consortium partners on the final day of the Summer School. It consists of guidelines for further iterations of the learning programme.

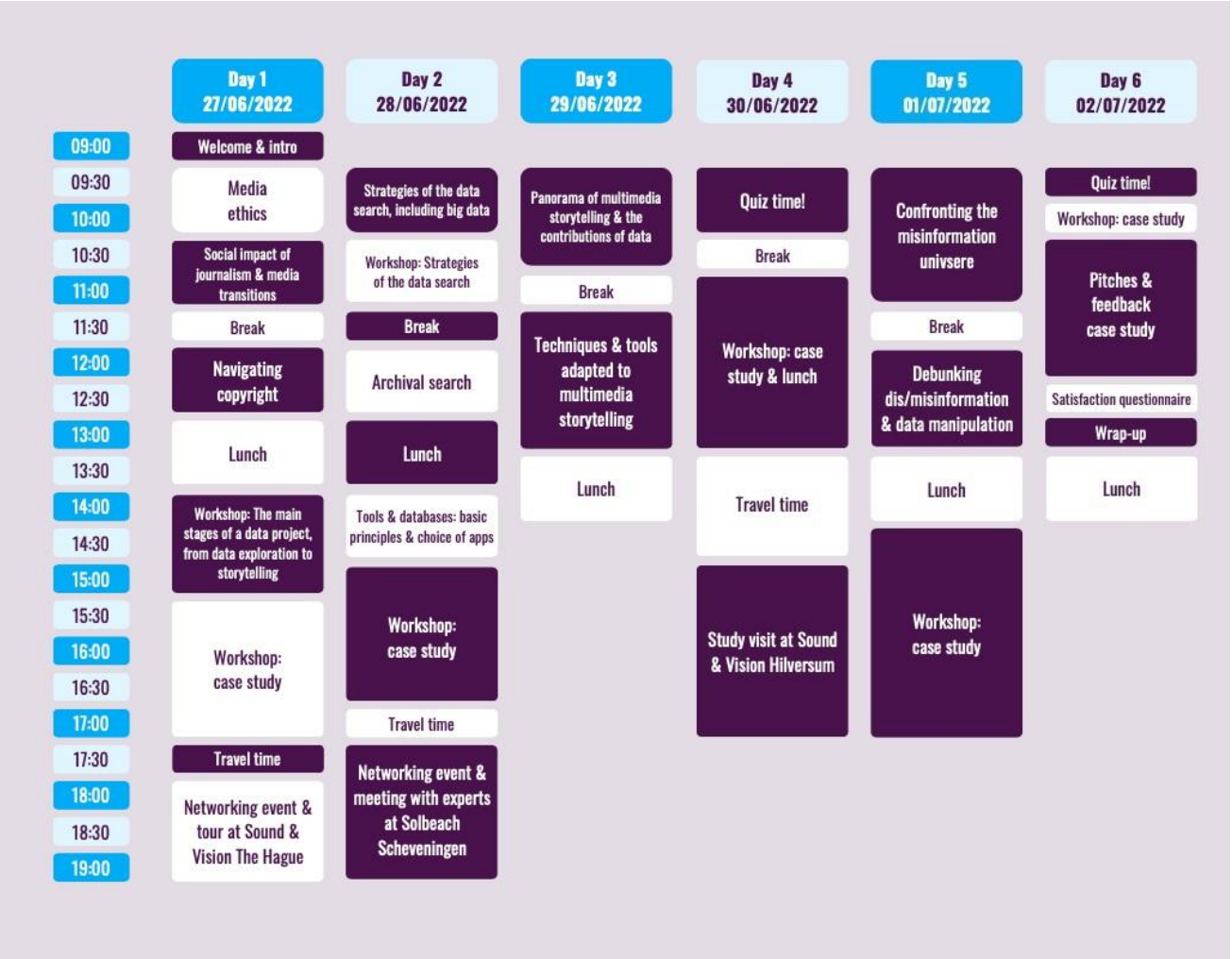


Figure 1. Detailed programme of the Summer School.

3. Detailed Syllabus

The final syllabus was constructed as an iterative process, which started in mid-2020 (in M6) by assessing which parts of the entire curriculum will be useful and how they can be incorporated into the programme. Crucial parts of the syllabus creation were also the Needs Analysis and the State of the Art reports. The programme was based on findings provided by extensive desk and field research. Hence, assumptions made in the project description about the topics were not entirely supported by the needs assessments' results. After conducting these analysis, we made several decisions. Firstly, we were determined to focus on students with minimal or basic data skills. Additionally, we actively sought out teachers who were comfortable instructing in such an environment.The initial programme description was juxtaposed with these findings, resulting in a programme reflecting these findings. As a result, a set of three syllabi was created, one for each module, and were then merged into one coherent document. The below section covers all of the contents of the syllabus, but if you would like to see the actual syllabus document, you can view it here.

3.1. Brief Description of the Course

The course aims at providing the students with basic ways in which data can be used in journalism, with concrete examples of tools used in data gathering, analysis and visualisation. Since part of the process is making sure no misinformation is spread, part of the curriculum will be devoted to strategies for combating it. The programme is divided into three modules.

The first module, named “Search and exploration of web data”, revolves around real-life cases that use small-scale data sets, and explains the process of finding, interpreting and using data. Its goal is not only to introduce students to working with data but to inspire them to continue to do so. It is based on the assumption that basic journalistic principles/storytelling skills are essential to data journalism and data-driven storytelling. The course provides the basis for acquiring basic data literacy skills and helps overcome the concerns created by numbers and technology.

The third training module “Telling stories with multimedia data” consists of providing the participants with the basic skills and know-how to be able to tell stories with data. This module takes the form of face-to-face courses delivered by expert trainers in the field, combining theoretical contributions in the classroom, feedback from professionals and a workshop, to put the teachings into practice.

The last module, “Tracking and debunking misinformation” aims at providing the students with an overview of how fake news fits in a broader reflection about media, describes examples of methods used in debunking them and works on a framework that can be used in the students’ further projects.

The programme is closely intertwined with a case study that the students solve in small groups that will put the above ideas into practice. The course concludes with a study visit at the premises of a project partner.

3.1.1. Objectives of the Course and Learning Goals

The following outlines the objectives and learning goals of the MediaNumeric training course:

- Students will be able to apply basic skills in finding a story and information to data-driven storytelling;
- Students will be able to find relevant data sets and extract the relevant information for a story;
- Students will be able to collaborate with professionals from other fields and establish 'common grounds' while creating a data-driven story;
- Students will understand the basic principles of human-computer interaction/computational thinking;
- Students will be able to find and extract relevant data from digital databases (data capturing skills);
- Students will understand the concept of storytelling and its evolution in the digital age;
- Students will understand the implication of data in storytelling and its recent uses;
- Students will be able to read, understand and interpret basic statistics, as well as use basic statistical tools (ie. Google Sheets);
- Students will understand the potential of the most common data visualisation tools;
- Students will be able to implement basic data storytelling projects;
- Students will have general knowledge of the contemporary media ecosystem, its role for democracy and challenges it faces;
- Students will have specific knowledge on misinformation in media, including different kinds of fake news, media specificity, and emergence of new technologies and tools to disseminate them;
- Students will have specific knowledge on the social obligations of the journalism profession, its responsibilities and ethics;
- Students will become acquainted with the contemporary journalist's toolbox and procedures to organise their work;
- Students will be able to spot misinformation and distinguish it from facts;
- Students will be able to use a framework to debunk fake news;
- Students will develop a more critical attitude, become better informed and more responsible professionals and citizens.

3.1.2. Prerequisites, Admission Requirements & Procedures

Students involved in the MediaNumeric training course are expected to meet the following criteria:

- English proficiency at least at B2 level with the capacity to read texts in English and participate in group discussions;
- Complete pre-work and reading assignments prior to the course (the students will be given the full list of materials to get acquainted with during the Day 0 introduction);
- Participate in the Day 0 introductory session.

3.1.3. Class Rules, Assessment Criteria & Methods

Students involved in the MediaNumeric training course are expected to adhere to the following:

- The students will be asked to work in small groups (4-5 people) around a case study presented during the Day 0 session. The task is to work on the case and present the results to the entire cohort and tutors at the end of the week;
- The students are expected to attend all classes and activities in person and bring their own laptops (if possible);
- There will also be a set of small analytical tasks to check the students' understanding of the contents covered throughout the course. At the end of each module, participants will be assessed to ensure that the knowledge has been acquired. These evaluations take place in the form of a quiz with 30 questions (1 point per question), to be completed over a period of 30 minutes. Participants who obtain an average of 50% correct answers on the quiz - or 45 points out of the 90 points in total - will pass the training course and obtain a certification.

3.2. Course Contents

This part consists of detailed information covering each of the three modules, outlining details of each course.

3.2.1. Day 0 Online Introductory Session - Pre Summer School Class

Day 0 took place on June 16, 2022, eleven days before the Summer School in The Hague. The goal of this online session was for the students to get to know the MediaNumeric team and each other, as well as get acquainted with the basic premise of the learning week. The students had a chance to listen to the results of the *State of the Art* report, participated in a short workshop in using Chartmetric tool, and heard about the idea behind the programme, including the case study, which was introduced during the meeting.

The three-hour meeting consisted of the following sections:

1. Introduction to the MediaNumeric Programme
2. Introduction to the topic of data journalism: State of the Art report
3. Introduction to Chartmetric

4. Introduction to the case study
5. Wrap-up and Q&A session

3.2.2. Summer School in The Hague

The Summer School took place between 27 June and 2 July 2022. The MediaNumeric team, accompanied by the teachers, led the programme throughout six days, detailed below.

3.2.1.1. Module 1: Search & Exploration of Multimedia Data

The below paragraph details the topics and related content covered across Module 1: Search & Exploration of Multimedia Data.

Navigating Copyright

Copyright is a pervasive aspect of digital media. This workshop provides a level playing field with key concepts of copyright within the European Union and specific aspects of copyright that are specific for journalism, like:

- Citing sources
- Text- & Data-Mining
- Adding media

Archival Search

In this course we dive into the broadcasters' catalogue of the Netherlands Institute for Sound & Vision and see how the following aspects can influence archival search and exploration possibilities:

a. Different metadata generation methods, e.g. fetching production metadata and subtitles, manual annotation, automatic metadata generation;
b. Metadata modelling;
c. Structuring (or not structuring) of metadata (with or without a thesaurus);
d. Shifts in selection and annotation policies;
e. Metadata incompleteness, distortions and biases inherent to different workflows.

Strategies of Data Search

a. Why data?
b. Where to find data
c. Data/tool criticism
d. Spotting patterns
e. Putting data into context
    i. Need for research question
    ii. Need for domain knowledge

Know your data. Know your tools. Know your domain

Stages of a Data Project

a. From ideas to data
    - Motivation
    - Resources
    - Collect
    - Aim + Context = Research Question
b. From data to stories
    - Tidy up
    - Understand
    - Communicate
c. Data workflow: programming vs. clicking -> advantages and disadvantages of both approaches
d. Science vs. Journalism

Tools & Databases: Basic Principles & Choice of Apps

We will briefly discuss the types of data that are available online and the tools to access and retrieve them. Particular emphasis will be placed on the use of these tools and their impact on the analysis that can be made of the data thus obtained.

a. What kind of data is needed?
b. Potential biases in the data pipeline (creation, distribution & harvesting)
c. Data formats (unstructured, semi-structured, structured)
d. Data sources (open data, dump, API, HTML, manual)
e. Tech / legal issues
f. Tools & examples

3.2.2.2. Module 2: Telling Stories with Data

The below paragraph details the topics and related content covered across Module 2: Telling Stories with Data.

Panorama of Multimedia Storytelling & the Contributions of Data: State of the Art of Storytelling Multimedia. Why Tell a Story with Data?

a. Why and how do we use data in telling a story?
b. Introduction: why tell a story with data? Data and data visualisation as a rhetorical tool to demonstrate and illustrate a point
c. What is storytelling with data? Based on storytelling = why and how

Techniques & Tools Adapted to Multimedia Storytelling: The Different Techniques for Designing a Data Storytelling Project

a. The actors involved: journalist/copywriter/author, data analyst, data designer, designer/illustrator, photographs, developer, motion designer, producer, etc
b. The tools involved: pen and paper, world, excel, coding for data analysis (R, Python, D3, etc), design software, visualisation software, coding for web (Python, HTML, js, etc), animation software, etc
c. Data visualisation tools: free data visualisation software (Flourish, rawgraphs, Kartis, Tableau, etc)
d. Introduction to Flourish

3.2.2.3. Module 3: Tracking & Debunking Misinformation

The below paragraph details the topics and related content covered across Module 3: Tracking & Debunking Misinformation.

Media Ethics

a. Introduction to ethical challenges of journalism
b. Legal regulations, professional ethics and codes of conduct
c. Problematic cases and examples showcasing different ways in which media ethics guidelines were breached
d. New forms of unethical practices connected to digital media and data manipulation
e. Ethical decision-making models
f. Big data, data visualisation and its risks

Confronting the misinformation universe

a. The misinformation universe: definitions, origins, dynamics and environment Anatomy and itinerary of false information: case study
b. Data and science in a fast-changing world: popular examples of misinterpretation of data
c. Why do we fall for false or misleading information
d. Basic rules of verification: The journalist's 5W toolkit
e. Out of context: how to debunk photos and videos shared on social media

Fake News & Digital Manipulation / Open Source Intelligence

a. Digital Manipulation: How to discover and verify a photo montage, a doctored video, or tinkered sound in social media content
b. Geolocation: How to find any place on Earth and why is it important in fact-checking (Suncalc, Google Earth)
c. Advanced Search: Google + Twitter (Crowdtangle)
d. Archiving as an Accountability Tool: How to fight cowardice and prevent link rot

e. Deepfakes Phenomenon - Are they a future of disinformation or are they irrelevant? Tools and materials (link), Deep Fake Detection Challenge overview (link, link)
f. Monitoring: Creating lists of the usual suspects (Crowdtangle, BuzzSumo)
g. Underbelly of the Internet: Dummy profiles, Infiltrating closed communities and learning the language of the "other" tribe (Closed FB groups, Telegram, Whatsapp)
h. Other tools for Fighting Disinformation: InVID plugin, MapChecking, and other open-data sources

Social Impact of Journalism & Media Transitions

a. Consequences of the internets' platformization and new business models on media ecosystem and democracy
b. Democratisation of the access to the audience, especially social media - new media practises and experiences (Media formats mix)
c. Internet fraud, social antagonisms, depreciation, false positive beliefs, algorithmic inclusion/exclusion
d. Automated journalism (impact on the news)
e. Social consequences of new media models
f. User-generated content and its influence on media streams

3.3. Readings

The students were asked to read the two texts listed below as required reading for the course. Other reading materials, listed under the "Suggested Reading" section, were meant to support the classes and were suggested to the students who wished to further their knowledge.

3.3.1. Pre-Reading (Required)

- Dario Compagno, 2016, Families of practices. A bottom-up approach to differentiate how French candidates made use of Twitter during the 2014 European Campaign, in: "Tweets from the Campaign Trail: Researching Candidates' Use of Twitter During the European Parliamentary Elections" (Alex Frame et al., ed), pp. 33-52.
- Jacqueline Pietsch and Daniel Sorabji, 2021, State of the Art report in storytelling with data and misinformation.

3.3.2. Suggested Reading

Suggested Reading in Module 1: Search & Exploration of Data

- Ashley Fell, 2017, Why storytelling is so powerful in the digital era, in: YouTube/TED.
- Lisa Raehsler, 2013, What people search for – most popular keywords, in: SearchEngineWatch.com.
- Jacquelyn Bulao, 2022, How Much Data Is Created Every Day in 2022?, in: Techjury.net.

- Brahim Zarouali, 2020, Persuasion Effects of Psychometric Targeting and Chatbots, in: Social Media & Politics.
- Chantel Ridsdale, James Rothwell and Mike Smit, 2015, Strategies and best practices for data literacy education: Knowledge synthesis report.
- Guillaume Plique, 2020, Empowering social scientists with web mining tools, in: FOSDEM 2020.
- Facebook offers a distorted view of American news, in: The Economist.
- Facebook's Top 10, a Twitter account.

Additional Reading in Module 1: Search & Exploration of Data

- C. Dawson, C., 2020, A–Z of Digital Research Methods. Routledge, Chapter 23.
- N. H. Riche, C.Hurter, N. Diakopoulos and S. Carpendale (Eds.), 2018, Data-driven storytelling. CRC Press, pp. 59-85.
- A. Khan and H. Din, 2021, New Age Journalism and Big Data (Understanding Big Data and Its Influence on Journalism). Big Data Analytics for Internet of Things, pp. 333-349.
- S. Knowles, 2018, Narrative by Numbers: How to Tell Powerful & Purposeful Stories with Data. Routledge. pp. 1-67.
- I. Foster, R. Ghani, R. S. Jarmin, 2020, Big Data and Social Science, 2nd Edition. Routledge. pp. 23-34.
- B. Franklin, S. Eldridge II (Eds.), 2016, The Routledge companion to digital journalism studies. Taylor & Francis, (Part V)**.**
- R. Rogers, 2019, Doing digital methods. Sage, pp. 1 - 41.

Suggested reading in Module 2: Telling Stories with Data

Visual Storytelling

- Life in the Camps and Tracking China's Muslim Gulag, by Reuters Investigates.
- Lost in Europe.
- The Pudding.

Dataviz

- Graphs & tabs
  - What's going on in this graph and The New York Times' front page of March 27th 2020, by the New York Times.
  - Iraq's Bloody Toll, by South China Morning Post.
- Numbers without graphs
  - The Twitter Presidency, by the NY Times
  - One Angry Bird, by Periscopic.
- Data Visualisation state of the art
  - The U.S. interactive election maps, by ABC News.
  - Out of sight, out of mind.
  - What's my place in the world population?
  - How much warmer is your city?, by the BBC.
- Data Visualisation
  - Can Visualization Elicit Empathy? Our Experiments with "Anthropographics".

  - Connecting with the Dots.
- Pragmatism
  - Collected Papers 5.9, by Charles Sanders Peirce.

Suggested Reading in Module 3: Tracking & Debunking Misinformation

- S. Zuboff, 2018, The Age of Surveillance Capitalism, Profile Books.
- J. van Dijck, T. Poell, M. de Waal, 2018, The Platform Society: Public Values in a Connective World, Oxford University Press.
- T. Gillespie, 2014, The Relevance of Algorithms, in: Media Technologies, ed. T. Gillespie, P. Boczkowski, K. Foot, MIT Press.
- W. Chun, 2008, On "Sourcery", or Code as Fetish, "Configurations", vol. 16, no. 3.
- W. Chun, 2021, Discriminating Data, MIT Press.
- K. Crawford, 2021, Atlas of AI, Yale University Press.
- Jacob Soll, 2016, The Long and Brutal History of Fake News, in: Politico.
- Francesca Tripodi, 2018, Searching for Alternative Facts. Analyzing Scriptural Inference in Conservative News Practices, in: Data & Society, pp. 18 - 48.
- Bruce Schneier, Allie Wong, Samantha North and Mick West, 2021, The Battle for Truth: Disinformation, Misinformation, & Conspiracies, in: The CyberWire.
- The commitments of the code of principle, in: IFCN Code of Principles.
- Kristen Panthagani, 2022, 10 logical fallacies used in vaccine arguments, in: You can know things.
- M. West, 2020, Escaping the Rabbit Hole, How to Debunk Conspiracy Theories Using Facts, Logic, and Respect, Chapter 4.

Additional Reading in Module 3: Tracking & Debunking Misinformation

- Niam Yaraghi, 2019, How should social media platforms combat misinformation and hate speech?.
- Craig Silverman (ed.), Verification Handbook A Definitive Guide To Verifying Digital Content For Emergency Coverage, in: DataJournalism.com.
- Tommy Carl-Gustav Linden, 2017, Algorithms for journalism: The future of news work, in: The Journal of Media Innovations.
- Beate Josephi, 2016, Digital Journalism and Democracy, in: Digital Journalism, Sage, pp. 9-24.
- A. K. Schapals, A. Bruns and B McNair (eds.), 2019, Digitising Democracy, Routledge.
- Reuters' report on deep fakes, in: Reuters.
- Jonathan Albright, 2017, FakeTube: AI-Generated News on YouTube, in: Medium.
- Riana Pfefferkorn, 2020, Deepfakes in the Courtroom, in: Boston University Public Interest Law Journal. Vol. 29 Issue 2, pp. 245-276.
- Oscar Schwartz, 2018, You thought fake news was bad? Deep fakes are where truth goes to die, in: The Guardian.
- Lane Wilkinson, 2016, Post-truth, propaganda, and bullshit: a glossary, Sense and Reference, in: Sense and Reference.
- Giselle Rampersad and Turki Althiyabi, 2019, Fake news: Acceptance by demographics and culture on social media, Journal of Information Technology & Politics, 17:1, pp. 1-11.

- Cristian Vaccari and Andrew Chadwick, 2020, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News, in: Social Media + Society. January 2020.
- Julia Angwin (ed.), 2022, The Markup Splitscreen project, in: The Markup.
- Kevin Roose, Rabbit Hole podcast, in: The New York Times.
- Julia Bayer and Ruben Bouwmeester, 2022, Fact check: How do I spot a deep fake?, in: Deutsche Welle.
- AFP, AFP Code of Ethical Standards, in: AFP.
- Reuters, Reuters Standards and Values, in: Reuters.
- Brian Walski, 2003, Bronx Documentary Center, Altered Images, in: Altered Images BDC.
- Council of Europe Resolution on the Ethics of Journalism (1994), in: Accountable Journalism.
- Jean-François Furnémont and Tanja Kerševan Smokvina, 2017, European Co-Regulation Practices in the Media, in: The Council of Europe.
- Ethical journalism practices on migrants and refugees, in: European Journalists.
- Georgia Wells, Jeff Horwitz and Deepa Seetharaman, 2021, Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show, in: The Wall Street Journal.
- Damien Gayle, 2021, Facebook aware of Instagram's harmful effect on teenage girls, leak reveals, in: The Guardian.
- The Open University, 2017, The Potter Box, in: OpenEdu.
- Danah Boyd and Kate Crawford, 2011, Six Provocations for Big Data, in: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.
- Catherine D'Ignazio and Lauren F. Klein, 2020, Data Feminism, The MIT Press.
- Michael Schermann, 2019, A Reader on Data Visualisation, in; Github.
- Peter Haferl, 2019, The Ethics of Media Visualisation, in: Medium.
- Vladan Joler, 2021, New Extractivism, in: OpenSecret.

4. Course Materials

Below provides a list of all materials used during the Summer School classes. Each graphic is an interactive link that leads to the file of the presentation online.

4.1. Background Materials

4.1.1. Master Deck: Day 0

This document is a presentation used during an introductory meeting held online on June 16, 2022. The meeting aimed to introduce the students to the MediaNumeric team, as well as present them with details of the Winter School's programme.

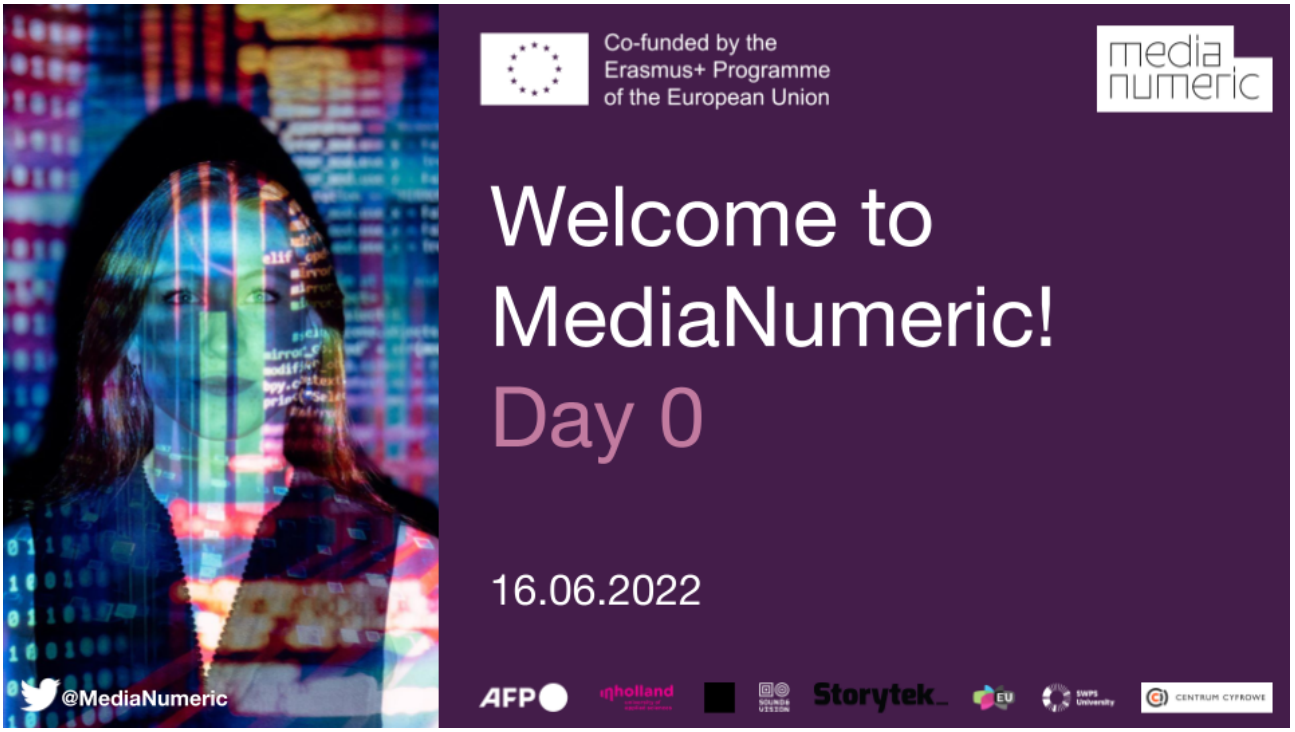


Figure 2: Day 0 slide deck.

4.1.2. Master Deck: The Hague Summer School

The master deck consists of slides used during the Summer School as backup material, with all necessary information, ie. the programme for a given day or planned breaks.

Figure 3. MediaNumeric Summer School Master Deck.

4.2. Lectures

4.2.1. Media Ethics (Lecturer: Mirek Filiciak)

The Media Ethics course is devoted to ethical dilemmas related to journalistic work - with tools to solve them and examples of good practices. The focus of the class is aimed at challenges that have arisen with the transformations of the media ecosystem, ranging from the functioning of classic problems in the environment of blurred categories, and ending with specific issues related to Big Data and data visualisation.

Figure 4. MediaNumeric Summer School, Media ethics.

4.2.2. Social Impact of Journalism and Media Transitions (Lecturer: Mirek Filiciak)

The Social impact of journalism and media transitions course systematises knowledge about the social consequences of Internet proliferation. Its theme is the emergence of new media structures and forms, but also the appearance of new intermediaries in the distribution of information, with growing importance of the technological component, which changed the way the public sphere functions.

Figure 5. MediaNumeric Summer School, Social impact of journalism and media transitions.

4.2.3 Navigating Copyright (Lecturer: Maarten Zeinstra)

Copyright is a pervasive aspect of digital media. This workshop provides a level playing field with key concepts of copyright within the European Union and specific aspects of copyright that are specific for journalism, like citing sources, text- & data-mining and adding media.

Figure 6. MediaNumeric Summer School, Navigating copyright.

4.2.4. Workshop: The Main Stages of a Data Project, from Data Exploration to Storytelling (Lecturer: Dario Compagno)

In order to produce a data analysis one has to move through several steps. Firstly, one needs to understand the context of the analysis: what are the needs grounding it? How can these needs be translated into research questions, and then into data collection? This requires intuition and clear ideas about our aims and means. Once we have this sorted out, the practical work begins. Data has to be treated for visualisation and modelling. It's here that software knowledge becomes central. In the end, our work has to be brought to the relevant public, so graphical and writing skills should now lead the way. This module will try to identify the most important points to check at each of these steps.

Figure 7. MediaNumeric Summer School, Strategies of a data project.

4.2.5 Strategies of data search, including big data (Lecturer: Mari Wigham)

In Strategies of Data Search, we will look at where (big) data can be found and how they can give us new insights. Crucial to this process are data/tool criticism and context, to ensure we discover real information from the data, and aren't fooled by data 'tricks of the light.'

⇒ Exercise data handout for the lecture here

Figure 8. MediaNumeric Summer School, Strategies of Data Search.

4.2.6 Archival Search (Lecturer: Tim Manders)

The Netherlands Institute for Sound & Vision collects an enormous amount of Dutch media heritage. In order to be able to process the amounts of media, and to keep them findable via metadata, over the years different annotation policies and practices were used. Currently with a focus on more automated metadata processing and the use of AI technologies like speaker voice detection and face recognition versus differentiated depth of descriptions per genre in the previous era of manual annotation. Obviously these changes have consequences for the search and retrieval of collection items. In this course we will dive into the broadcasters' catalogue of Sound & Vision and see how aspects such as metadata modelling, structuring (or not structuring) of data, selection and annotation policies influence archival search and exploration possibilities. Furthermore, we will take a glimpse at how different practices lead to different kinds of metadata incompleteness, distortions and biases that affect findability.

⇒ Handouts and additional reading for the lecture here

Figure 9. MediaNumeric Summer School, Archival search.

4.2.7. Tools & databases: basic principles & choice of apps (Lecturer: Nicolas Hervé)

We will briefly discuss the types of data that are available online and the tools to access and retrieve them. Particular emphasis will be placed on the use of these tools and their impact on the analysis that can be made of the data thus obtained.

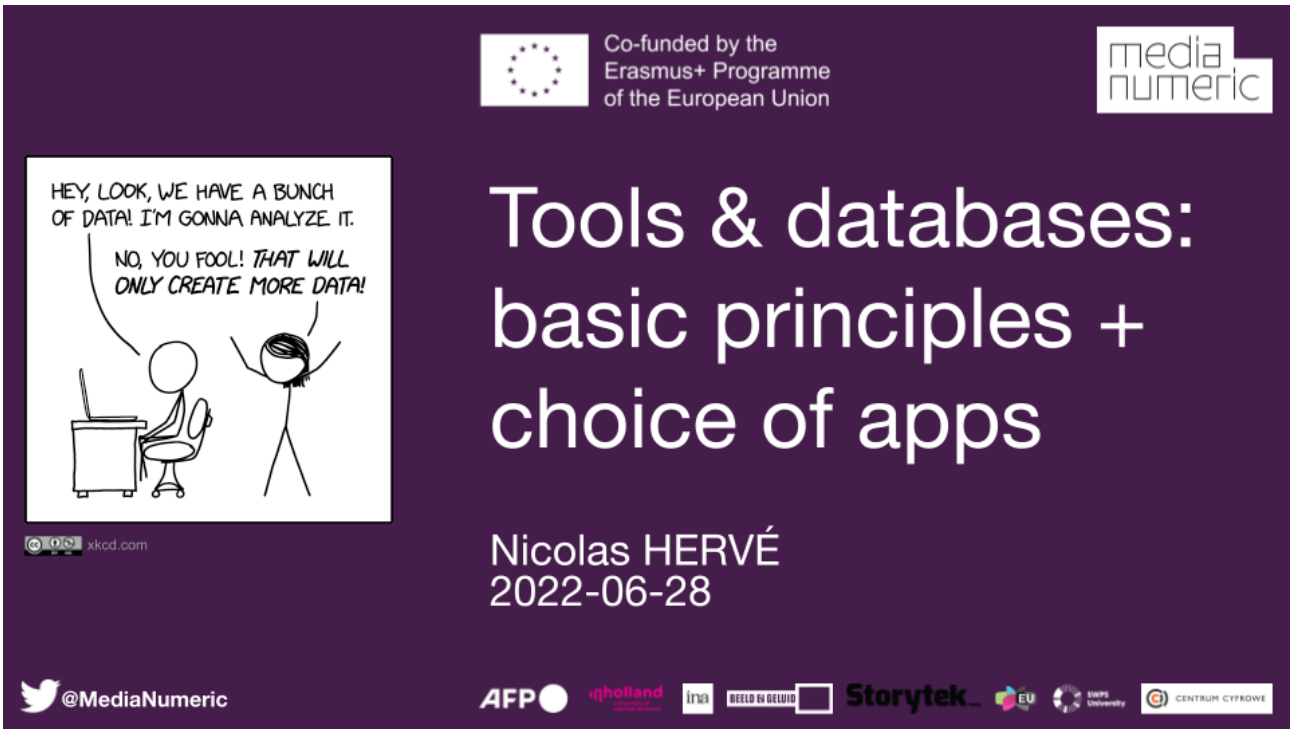


Figure 10. MediaNumeric Summer School, Tools & databases: basic principles + choice of apps.

4.2.8 Panorama of multimedia storytelling & the contributions of data (Lecturer: Julie Brunet)

In this lecture, we'll consider all the actors involved in a data storytelling project and their mutual tools before focusing on Flourish, a free online data visualisation tool.

⇒ Handouts and additional reading for the lecture here

Figure 11. MediaNumeric Summer School, Telling stories with data.

4.2.9. Storytelling with Data: Tools (Lecturer: Julie Brunet)

In this lecture, the students consider all of the actors involved in a data storytelling project and their mutual tools before focusing on Flourish, a free online data visualisation tool. The students get the chance to put their knowledge into practice using the aforementioned tool and data sets they started gathering to complete their case study assignment.

⇒ Lecture outline here

Figure 12. MediaNumeric Summer School, Storytelling with data: tools.

4.2.10. Confronting the misinformation universe (Lecturer: Robert Barca)

What is disinformation and misinformation? How do you recognize it? Where does it come from? How does it circulate, particularly on social media? Students will gain an understanding of a trend that has been identified as one of the biggest threats to democracy. They will learn where to start the process of verification and the basic tools to debunk manipulated content.

⇒ Fact-checking exercises here

Figure 13. MediaNumeric Summer School, Confronting the misinformation universe.

4.2.11. Debunking dis/misinformation & data manipulation (Lecturer: Robert Barca)

A deep dive into open-source tools that can be used to investigate, track, and debunk dis/misinformation and digital manipulation. Students will learn advanced search criteria, geolocation, image analysis tools and other techniques to identify manipulated content. They will also learn the importance of accountability and how to record their investigation so that their work stands up to scrutiny.

⇒ Fact-checking exercises here

Figure 14. MediaNumeric Summer School, Debunking dis/misinformation and digital manipulation.

4.3. The Case Study and the Pitching Session Materials

4.3.1. The Case Study

Below is a detailed description of the case study that the student groups completed. It was presented during the Day 0 online meeting and was a reference point for the students in their work on their pitches. The case study document is available here.

PICTURE: Amsterdam Dance Event
https://www.amsterdam-dance-event.nl/en/news/first-results-of-the-exclusive-research-into-diversity-and-inclusion-within-the-electronic-music-industry/164769/

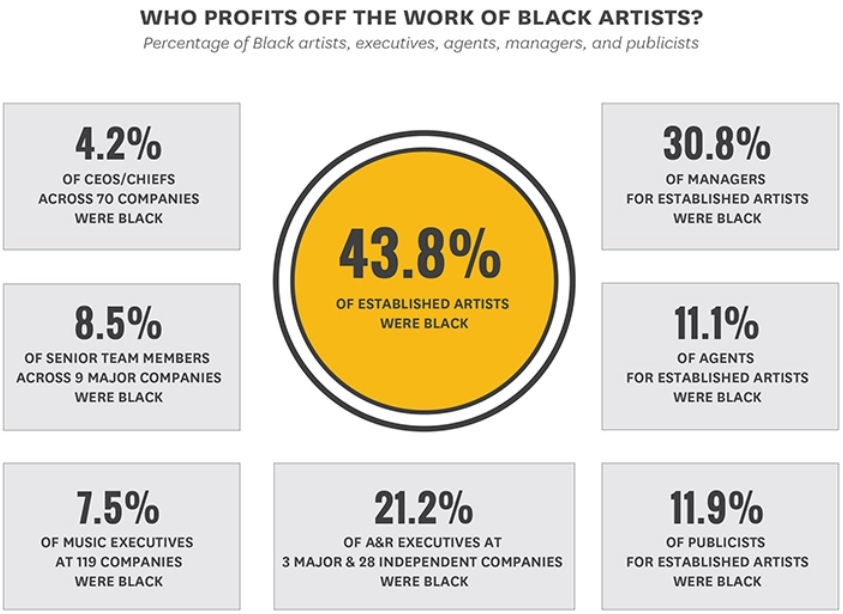
Context

Where is the music industry on diversity and inclusion? The international music industry — built on a foundation of love for music created by artists of all races, ethnicities, gender identities, and sexual orientations — is uniquely poised to help make a truly diverse and inclusive reality, if we choose to do so.

Inequalities based on race have been particularly important in shaping musical history. For example, Sonnet (2020) notes that the state of Mississippi offers an interesting case study into understanding both race and music in the USA. A historically important location of racialized slavery, segregation, and civil rights struggles, Mississippi also proclaims itself the 'Birthplace of America's Music' on welcome signs. Mississippi is a critical site for studying music and society, where the colonization of African American music meets the creativity and placemaking of present-day African Americans.

But how diverse is the music business? The answer, according to a new study: not very. A report by the Annenberg Inclusion Initiative, released in 2021, examined the makeup of 4,060 executives

at 119 companies, such as corporate music groups, record labels, music publishers, radio broadcasters, streaming services, and live music companies. In the 70 major and independent music companies, only 13.9% of top executives were from underrepresented racial or ethnic groups.



SOURCE: Data Visualization from Variety Magazine https://variety.com/2021/music/news/music-industry-diversity-usc-annenberg-report-1234996163/

Discrimination is not simply relegated to the boardroom, but can **also plague the algorithms that choose the music we consume.** Spotify's algorithmic delivery was what initially set it apart from other music streaming platforms, often cited as an important factor in the app's success despite how it relies on tracking data. While tracking music data does not immediately seem murky, the use of artificial intelligence has been proven to discriminate. Reports have shown how artificial intelligence can be encoded with bias and perpetuate racism.

With movements like #Metoo, **the barriers that women face in the music industry are becoming more acknowledged by the public**. Nevertheless, there are still many challenges and obstacles, many of which remain unsolved within this influential industry. For instance, in December 2020, Palestinian techno artist Sama' Abdulhadi was imprisoned for eight days after performing a live stream event at a historical site in Jerusalem. Although all permits were correctly secured, an angry crowd turned up in protest. Abdulhadi had intended to use the live stream to showcase artists from Palestine and the Middle East during the pandemic.

Gender issues today are by no means limited to women in the industry. They extend to people of color and anyone who is non-heterosexual or non-cisgender. These issues have moved further into an intersectional identity issue. This is why this year, the BRITS decided to remove gender-specific categories to be a more inclusive award show.

Choosing Your Datasets

The goal of this case study is to find and fact check data to **create a compelling story about inclusion and diversity in the music industry**. You can choose to take a more historical focus, a more technical focus, a more sociological focus, or a focus on current events. You are encouraged to be as specific as possible with your storytelling. Try to think like a journalist and select a focus that is relevant, recent, and interesting to you. If you are feeling stuck on how to begin, you can check out some sample research topics to consider (here).

Chartmetric

All groups are required to **use at least one raw data set** from our industry partner, Chartmetric. Chartmetric is a music data analytics tool. They collect public data from sources such as Spotify, YouTube, and TikTok and create datasets about artists, listeners, followers, social media engagement, and more. This data can be used to design convincing visualisations and identify powerful insights that help music industry professionals and journalists alike to tell impactful stories.

Other Datasets + Sources

In addition to Chartmetric, you are free to use any data set that your group agrees on. You can scrape the data from the internet, download them from any source, or use links to data sets and resources provided by the MediaNumeric team (here).

Tasks

As a group, you are expected to deliver the following tasks during the summer school experience between June 27 to July 2, 2022:

- **Develop a key question** your group wishes to work on - it's up to you how you approach the case and how you want to craft the message of your story
- **Gather data and include justification** behind picking particular data sets
- **Analyse** the chosen data sets and draw conclusions
- **Visualise** the data (if relevant to your chosen story medium)
- **Create a story** (approximately 1-2 pages) for a medium of your choice
- **Share** your work with the MediaNumeric team on **Saturday, July 2nd by 9:00 AM** via Participants' Google Drive (link). Please check the Deliverables list for what we expect to be uploaded.
- **Present the final data journalism story** in public to a panel of judges and an audience of your fellow participants (Saturday morning, July 2nd). You can present your story with a PowerPoint, video, or other media format.

Deliverables

Materials to Upload to the Drive

When sharing your final story onto the Google Drive, please share the following:

- a PDF version of your final data journalism story (or a relevant link to listen to/ watch your story)
- a folder with the dataset(s) used for the story
- a bibliography of sources used to create your story

Pitch Rules & Requirements

The Pitching Session aims to encourage you to learn how to sell your idea, in a limited amount of time. The final pitch should only focus on the final data journalism story, **not** the background information (i.e. the process you took to arrive at this final story). The latter will come up during the post-pitch discussion. You should expect questions about your data gathering process, decision making, group work dynamics, etc.

Rules & Format:

- Each group pitches in front of an audience (panel of experts & classmates)
- Each group has **7 minutes** for their presentation
- Each presentation is followed by a **10-minute feedback session** with Q&As from the panel of experts and a **5-min open Q&A** from experts & classmates
- Each group is free to use its own slide template if you wish, or any format of the presentation is acceptable
- There´s no minimum / maximum number of slides that can be used
- You are free to organise the structure of your pitch as you wish

Things to Consider Including in Your Pitch:

- Catchy title/headline
- Key question/problem you're tackling
- (A) data point(s) with explanation
- A visualisation of the data as an image, video, or other media format
- So What? Why is this important? Why should we care?

Additional Materials

Sample Research Topics

The goal of this case study is to find and fact check data in order to **create a compelling story about inclusion and diversity in the music industry**. If you are feeling stuck on how to begin, here are some sample research topics to consider.

Fake Artists in the Music Industry

According to DN, Firefly Entertainment is doing a roaring trade in what some would call "fake artists" on Spotify. These are the now well-known pseudonymous artists on the streaming platform – **artists with no discernible online footprint** – whose music fills up many of Spotify's key mood and chillout playlists. Who are the rightful owners of this music? How do these new music business practices affect diversity on the DSPs (digital streaming platforms)?

- **Possible source:** An MBW reader just blew open the Spotify fake artists story. Here's what they have to say: https://www.musicbusinessworldwide.com/an-mbw-reader-just-blew-open-the-spotify-fake-artists-story-heres-what-they-have-to-say/
- **Possible Source:** Remember Spotify's fake artists? They're still going strong – and still attracting scandal: https://www.musicbusinessworldwide.com/remember-spotify-fake-artist-theyre-still-going-strong-and-still-attracting-scandal/

Conflict and the Ukrainian Music Industry

Since 2014, Ukraine has had a rapidly growing music industry, diverse in genres and rapidly growing on the international stage. Has the popularity of Ukrainian-created music come to a halt since the conflict began or perhaps has this conflict driven more interest in the artists that call this country home?

- **Possible Source:** Photos From Ukraine: ONUKA, Alyona Alyona & Andriy Khlyvnyuk & More Musicians on the Front Lines: https://www.billboard.com/photos/2022-ukraine-war-musicians-photos-1235069684/
- **Possible Source:** 'I'm Destroyed Inside': Ukraine's Flourishing Music Scene Faces an Existential Threat: https://www.rollingstone.com/music/music-features/ukraine-musicians-russia-invasion-1313403/
- **Possible Source:** Ukraine's Eurovision win shows us that despite arguments to the contrary, the contest has always been political: https://theconversation.com/ukraines-eurovision-win-shows-us-that-despite-arguments-to-the-contrary-the-contest-has-always-been-political-182767

Music creators and Gender Inequality in the Dutch Music Sector

According to Berkers, Smeulders, and Berghman (2019), Only 13% of the members of Buma/Stemra – the Dutch collecting society for songwriters, composers, and music producers – identify as female. Historically, how pervasive has inequality been in the Dutch music industry? Does inequality affect certain genres or certain professional roles within the music industry more than others? How has the #MeToo movement affected the Dutch music sector? Where (if at all) do we see progress toward more inclusion concerning gender in the Dutch Music industry?

- **Possible Source:** Music creators and gender inequality in the Dutch music sector: https://www.aup-online.com/docserver/fulltext/13883186/22/1/03_TVGEND2019.1_BERK.pdf?expires=1652887558&id=id&accname=guest&checksum=9D0528B02F6B7252BC5E1B7D25A88C9F

- **Possible Source:** The Netherlands Had Their Own Big #MeToo Scandal – The Start of a Change? https://mediummagazine.nl/the-netherlands-had-their-own-big-metoo-sandal-the-start-of-a-change/
- **Possible Source:** The working life of musicians: mapping the work activities and values of early-career pop musicians in the Dutch music industry https://www.tandfonline.com/doi/full/10.1080/17510694.2021.1899499

Suggestions for Data Sources

In this section, we have provided a few data sources that can be used for this case study **in addition to the datasets found on Chartmetric.** Please be aware that this list is not exhaustive and that you are encouraged to also find your own data sets online as well to support your selected case study.

- **Eurostat Datasets about Music and the Cultural Sector:** https://ec.europa.eu/eurostat/web/culture/data
- **Composer Diversity Database:** https://www.composerdiversity.com/database-instructions
- **Pronoun and Gender Dataset** (Make Music Equal by Chartmetric): https://makemusicequal.chartmetric.com/pronoun-gender-database
- **Spotify Datasets** (stored on AICrowd): https://research.atspotify.com/datasets/
- **DataSearch Google** (please be aware note all datasets are free from this source): https://datasetsearch.research.google.com/

Data Scraping Tips and Tools

In this section, we have provided a couple of tools that might help you to collect and analyse your datasets in a cleaner/deeper fashion:

- **Dataminer** (a web scraping tool that operates as a Google Chrome extension): https://dataminer.io/
- **Open Refine** (A tool for cleaning messy data): https://openrefine.org/

Data Visualization Tips and Tools

In this section, we have provided a few data visualisation tools and examples that might help you to visualise your data:

- **Tableau Public** (Data visualisation tool): https://public.tableau.com/s/
    - Check out the **Tableau Gallery** as well for inspiration: https://public.tableau.com/app/discover/viz-of-the-day
- **Flourish** (Beautiful data storytelling platform): https://flourish.studio/
- **Infogram** (browser-based infographics generator): https://infogram.com/

4.3.2. Pre-Work and Pre-Read Document

The document linked here consists of all information important for the students prior to attending the Summer School, including required reading, technical prerequisites and basic information about the case study preparation work.

Pre-technical Installation

- Please connect to the student Google Drive and ensure that you have access.
- Please ensure that you have the latest versions of Google Chrome or Firefox installed on your computers.
- Add the InVid Verification plugin to your browser. For Chrome, click here. For Firefox, download directly from the InVid page.
- Add the Wayback Machine plugin to your browser. For Chrome, click here. For Firefox, click here.
- Add the CrowdTangle plugin to Chrome by clicking here. This plugin does not exist on Firefox.

Case Study Pre-work

- Please **write a short answer** to the following question: *How would you go about finding data about diversity and inclusion in the music industry? Where would you look for the data? Why?*
- Using the questions above, **select one data set from Chartmetric and at least one other data source outside of Chartmetric. Write a short justification about *why* you selected them.** Please have these data sources ready for Day 1 of our Summer School.
- Find an **example of a visualisation** of the topic of diversity and inclusion in the music industry, from any angle you think is interesting, and describe things you like, and don't like about it. Have it ready for our Day 1 meeting.
- Familiarise yourself with 3 short lessons dedicated to **Google Sheets**:
    - Google Sheets: Scraping data from the internet
    - Google Sheets: Cleaning Data
    - Google Sheets: Visualizing Data

Pre-reading

1. Compagno - The main stages of a data project - Families of practices
2. State of the Art report in storytelling with data and misinformation (by Jacqueline Pietsch)

Pre-watching/listening

1. Watch this YouTube video called **"We tracked what happens after TikTok songs go viral"**

2. Watch this video on **storytelling with data**
3. If possible, watch ***Spotlight*****,** the 2015 film by Tom McCarthy about the true story of how the Boston Globe newspaper uncovered the massive scandal of child sex abuse within the local Roman Catholic Archdiocese, shaking the entire church to its core.

4.3.3. Pitching Session Support Deck

The deck below consists of the pitching session rules and was presented as a visual background for the session during which the students presented their work. The rules were:

- Each group will pitch on the stage in front of audience;
- Each group will have 7 minutes for their presentation;
- Each presentation will be followed by a 10-minute feedback session with Q&As from the panel of experts and a 5 minute open Q&A;
- Each group is free to use its own slides template - but you also do not have to - any format of the presentation is acceptable;
- There´s no minimum/maximum number of slides that can be used;
- You are free to organise the structure of your pitch as you wish.

Figure 15. Slide deck pitching session.

4.3.4. Student Presentations

On the last day of the Summer School, the students presented their work to the other groups and to the MediaNumeric consortium representatives. There were four groups, assembled prior to the Summer School by the MediaNumeric team. The groups were created in a way to assure gender

balance, but also a mix of students from different institutions and countries. However due to covid and other reasons only three groups pitched. Below are the descriptions of the stories created by the students. The presentations themselves are not shared to make sure the students' anonymity is preserved.

Group The Powerpuff Girls shared their presentation entitled "From a Microphone to a Feeding Bottle in the Backstage area: How Musicians' Lives Changed after Becoming Parents?" that aimed to understand the situation of parents in the music industry based on the publicly available data sets.

Group EXP shared their presentation entitled "A peek into American cancel culture," which provided a perspective on if cancel culture has any impact on the artist's career in the United States. The presentation was based on data from media and Chartmetrics data sets.

Group Team Matrix created a presentation "Creating a seat at the table" about the LGBT community and coming out in the music sector.

5. Study Visit

As part of the MediaNumeric programme, a study visit is planned to the offices of one of the consortium partners during each of the three training sessions. The aim is to demonstrate to the students how the skills learnt in the classroom can be applied in a real-life professional setting.

For The Hague school, the study visit was scheduled at The Netherlands Institute for Sound and Vision (NISV). NISV is one of the largest audiovisual archives in Europe with the collection of more than 750,000 hours of audiovisual content: television, radio, music and film. NISV serves as an archive for the largest national broadcasters, as well as a museum and cultural heritage institution.. NISV NISV also functions as a research institution for audiovisual content and actively participates in the creative reuse and reinterpretation of its collections.

5.1. Identifying the Goals of the Visit

In order to make the visit as relevant as possible to the course, the MediaNumeric project manager for NISV consulted with consortium members on the students' needs and expectations. It was decided to show the students the different stages of audiovisual content life: from creating, archiving, restoring, and finally making available online or at a venue for different purposes: artistic, scientific, educational and as a source of data for journalists. Once the needs were listed, the MediaNumeric project manager for NISV identified the relevant people within the archive who could best represent and relay these processes.

5.2. Preparing for the Study Visit

With social distancing restrictions still in place to limit the spread of Covid-19, the group was divided into two for safety reasons.

Three key people were identified, approached and secured to lead the presentations for the students: the research department manager, archival team members and researchers working on the NISV collection.

6. Quizzes and knowledge verification

In order for students to pass the MediaNumeric training course they were required to sit for quizzes that covered all lectures taught spanning the three main modules of the course. The students were required to take the quizzes in two parts. The first part consisted of quizzes covering Modules I and II, the second quiz referred to the material covered in Module III. The students were given 90 minutes in total. Below are all three quizzes listed with students' answers and a complete list of questions. Wording of the quiz' questions align with the quizzes that the students took.

⇒ Module 1 quiz
⇒ Module 2 quiz
⇒ Module 3 quiz

Quizzes are part of a broader evaluation process carried out mainly by INA, with support from other partners. The 'Evaluation Methodology' report refers to three stages of the evaluation:

1. "**Assessment of the level of knowledge and skills of the participants before the training course**: verification of the understanding of the contents covered throughout the course;
2. **Assessment of the knowledge and skills acquired by the participants at the end of the training course**: alignment of the results of the training programme with the pedagogical goals;

3. **Assessment of participants' satisfaction with the training course**: course content, pedagogical approach, formats and methodology, speakers, etc."

The quizzes are part of the second point in the above list which aims at assessing the students' knowledge and skills acquired during the Summer School and were designed to measure participants' progress in all three modules: "the examination of learning outcomes of the onsite training session consists of end-of-training validation tests (...). The MediaNumeric programme awards academic credits (5 ECTS for the on-site training programme (...), detailed in a transcript of records and skills. (...) The examination of learning outcomes for each on-site training programme will be conducted through end-of-training tests designed in the form of quizzes. The quizzes will be based on the teaching modules (...), the needs analysis (...), and the EQF (European Qualifications Framework)."

Based on these assumptions, the following quizzes were created and used as an evaluation tool during the Summer School.

6.1. Module 1 Quiz: Search and Exploration of Data

Correct answers are in bold.

Navigating Copyright

1. Freedom of the press means that as a journalist I can copy and reproduce everything I need for my article
   a) True, freedom of press is more important than copyright
   b) False, freedom of press is about the expression of ideas not reproduction of other people's work

2. I can use everything on the internet for free
   a) True, everything on the internet can be copied and reproduced in my own publications
   b) True, everything on the internet can be copied and reproduced in my own publications as long as I mention my sources
   c) False, I can only use certain things from the internet and only for certain specific purposes

3. Everything without the © sign is free from copyright
   a) True
   b) False

4. Is the exception to copyright for journalistic purposes the same in every EU country?
   a) Yes
   b) No

5. What is the duration of copyright in the European Union?
   a) Copyright ends when all the authors are dead
   b) Copyright ends when the latest living author died more than 70 years ago
   c) Copyright never ends
   d) Copyright ends 100 years after publication

6. Are you as a journalist allowed to download and store data to analyse data for a journalistic article?
   a) Yes, if they are publicly accessible and you can only store them for as long as they are needed for the research
   b) No, only universities can do that
   c) Yes, there are no limitations on analysing publicly accessible data

7. Who owns the right of an article that a journalist wrote for a newspaper?
   a) The newspaper, this is commonly accepted
   b) The journalist, if they made no other agreements
   c) Both

Strategies of data search

8. What do we mean by 'objective data'?
   a) It is not affected by people
   b) It is the truth
   c) It is measured in a consistent and verifiable manner
   d) It is reliable

9. What is data criticism?
   a) Critically evaluating the data you use
   b) Reporting data errors to a data provider
   c) A protest movement challenging the use of personal data
   d) Criticising the use of data in journalism

10. What is a unit test?
    a) a check if the units in data are correct
    b) a software test for an individual unit of code
    c) a check whether an organisational unit delivers open data
    d) a software test for unit conversions

11. What is tool criticism?
    a) quality control of tools during development
    b) opposition to the use of tools instead of programming yourself
    c) critically evaluating the tool you use for a given task
    d) crowdsourcing reviews of data tools

12. What is an outlier?
    a) a fake piece of data
    b) a data point differing significantly from the others
    c) a missing data point
    d) a data point excluded from further analysis

13. What is CSV?
    a) A mathematical software
    b) An American data journalism outlet
    c) A data file format
    d) A forum specialised in Data Queries

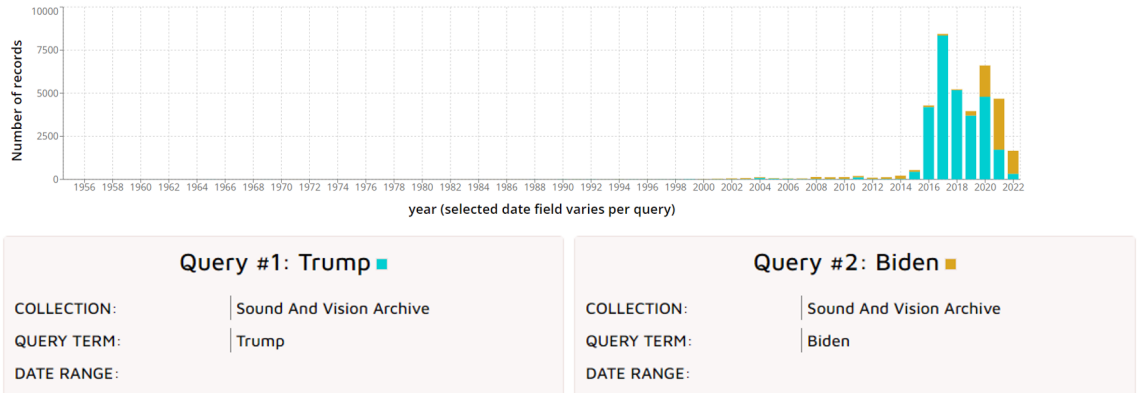
14. What is a Pivot Table?
    a) A tool to easily reshape your spreadsheet's data
    b) A feature that cleans your spreadsheet's data
    c) A filter that allows you to select the data you are looking for
    d) A cell's function

15. Which problem is **not** visible in this spreadsheet?

	A	B	C	D	E
1	**Music label**	**Artist**	**Gender**	**Number of records**	**Recording date**
2	Music NOW	A. Artist	Male	3	1st May 2022
3	Music NOW	Sid Singer	Male	2	4th June 2022
4	Music NOW	Gisela Guitaris	Female	3	8th August 2021
5	It's Music	B. Rapper		4	1st July 2020
6	It's Music	Fun Bob	Female	3	2-4-2020
7	Its Music	Sally Singer	Female	2	2nd December 2021
8	All Music	Artist, A.	Female	1003	1st July 2015
9	All Music	Ken Keyboard	Male	4	

    a) missing data
    b) odd data values
    c) erroneous data
    d) inconsistent naming

16. What does this graph tell us?



a) Donald Trump appears more frequently in the archive than Joe Biden
b) Donald Trump got more media attention during his presidency than Joe Biden
c) The search term 'trump' has more hits than the search term 'biden'
d) The search term 'trump' occurs more often than the search term 'biden'

17. Substantive expertise is essential when analysing data to:
a) correctly interpret results
b) check that data is correctly analysed
c) find out if data is trustworthy
d) write good code

Archival search

18. Which of the following metadata is most likely fully created manually after a media production is ingested into the archive?
a) Subtitles
b) A list of the cast and crew
c) A summary of the content in context
d) A person's name with a confidence score

19. What is not a true disadvantage of a hierarchical metadata model and search index?
a) Search performance can be slower because of hierarchical layering
b) Metadata for each new media production has to be generated for all hierarchical levels
c) Media formats, cast, crew and content change all the time so the top level metadata can only be very unspecific
d) Users can be confused about how metadata is distributed over different metadata levels

20. What's a thesaurus?
a) An autocomplete functionality containing all words in a database

b) Automatically created recommendations based on a users search history
c) A special type of dinosaur
d) A controlled and structured vocabulary with relationships between the labels

21. What is not a likely reason for a persons' names to be left out from a catalogue description?
a) The thesaurus is incomplete
b) They did not deserve it because their opinions are terrible, so they were consciously left out by the documentalist
c) Certain genres are traditionally annotated less extensively than others
d) The face recognition is based on a database with limited face models

22. What is the best way to search for a topic that was of interest in the media over a long time period?
a) Search with a generic term so you get back as much results as possible
b) Search with a very specific term so you get little but relevant results
c) Diversify your search strategies in accordance with changes in the discourse and archival workflows
d) Only search with a thesaurus label, because thesaurus labels are the only way to pinpoint all the content with that topic

23. Why do automatic metadata generation methods give biased results?
a) because AI training sets are not diverse
b) because people designing algorithms are often purposely discriminatory
c) because manual metadata is always better
d) because production metadata is always better

Tools & databases: basic principles + choice of apps

24. Your editor urgently asks you to write an article with a graph on the financial data of a company that is in the news because of a scandal. The data is in the form of a simple table on a web page. What is the best approach to respond as quickly as possible?
a) ask the software engineer to help you by creating a dedicated scraper
b) as you have learn the basics of python, write the scraper yourself
c) copy / paste the data in Excel and tidy it up
d) call the company and ask for a dump of the data

25. The HTML code of a web page can be accessed by:
a) double-clicking on the web page
b) using the developer tools of a web browser
c) logging into a web server using a special admin access
d) it is forbidden to access the HTML code of a web page

26. You need to do a study on the expression of politicians on Twitter over the last 5 years. You have a list of 100 accounts that a colleague has passed to you and you know how to query the Twitter API. What potential problem will you not face?
a) these accounts are not representative of your country's politicians

b) some tweets from the period have been deleted
c) your API access key does not allow you to retrieve tweets
d) you can't retrieve tweets that are too old

27. What is the limit of web scraping?
a) only very experienced developers (20+ years) are able to do it
b) you need a specific scraping licence only given by the government to a handful of developers
c) it is forbidden by the GDPR (General Data Protection Regulation)
d) some websites prevent scraping by blocking visitors with unusual behaviour (such as visiting too many web pages in a limited period of time)

28. You want to quickly automate the download of a hundred web pages, which tool allows you to download them?
a) firefox
b) jq
c) curl
d) htmlq

29. You find a csv file with information on arms sales to third world countries on a hidden web site claiming to be a NGO. What problem will you not have using this file?
a) this file format is not well structured
b) you will not be able to quote the author of the database
c) the completeness of the data is not guaranteed
d) the accuracy of the data is not guaranteed

30. If you have the choice between all these solutions to obtain the history of the stock exchange prices of the year 2000 in the USA, which one do you choose?
a) use the NYSE (New-York Stock Exchange) free API that allows to search for the share prices of each company
b) get the data already nicely formatted in a csv file by a student, that has done a thesis on a similar subject, provided on her website
c) ask by email to your colleague working in the USA
d) download a full dump from the NYSE web site

6.2. Module 2 "Telling stories with data"

Panorama of multimedia storytelling and the contributions of data. State of the art of storytelling multimedia. Why tell a story with data?

1. What phenomenon is called « infobesity »?
   a) a modern tendency of over-producing data
   b) a modern tendency of over-consuming data
   c) an ecological issue of storing all the data we create
   d) a personal cognitive incapacity of dealing with all the data we create

2. What main reason makes that a bare graph is usually not enough to engage an audience?
   a) it's too complicate
   b) it over-simplifies reality
   c) **it's not engaging enough**
   d) it's too long to understand

3. What main reason makes that a bare graph is usually not enough to create empathy?
   a) we have too much biases
   b) **we have issues understanding big amount of numbers, we need exemples**
   c) we're too much solicited in modern society
   d) we have trouble understanding it and its issue

4. What phenomenon is called « statistical numbing »?
   a) **the difficulty we have to process statistics and their implications in real life**
   b) the difficulty we have to remember statistics
   c) the difficulty we have to visualise statistics
   d) the difficulty we have to create statistics on certain events

5. In scrolly-telling, what do we call the trigger?
   a) the action of opening the internet explorer
   b) **the action of scrolling**
   c) the action of creating emotions through the story
   d) the action of sharing the story on social media

6. What is the most basic way to create a beginning of storytelling within a chart?
   a) adding a caption
   b) adding an illustration
   c) adding an explanatory title
   d) **adding an exploratory title**

7. What is an "outlier" in data?
   a) an external variable
   b) **a value extremely different from the rest of the dataset**
   c) an unexpected trend
   d) an added variable that changes the dataset

8. What is the main reason why design is a powerful tool in storytelling with data?
   a) **it convey emotions in an immediate way**
   b) it's understandable by everyone
   c) it take less time to read
   d) it's better memorise by people

9. What does Jake Harris mean by: we must « connect to the dots » in data visualisation?
   a) we must understand how data are connected to each other
   b) **we must see beyond data to connect with the real people behind it**
   c) we must connect to the audience through storytelling
   d) we must rely on data to tell stories

10. What steps or tools can you use to introduce more storytelling with data?

Having exploratory title, choosing the right chart for the right mental representation, using an incarnation of your data to start from, using scrollytelling, using interactivity, etc

Techniques and tools adapted to multimedia storytelling. The different techniques for designing a data storytelling project

11. Which one of these positions is usually NOT involved in storytelling?
   a) data analyst
   b) illustrator
   c) **data engineer**
   d) journalist

12. Which one of these tools is usually NOT involved in storytelling?
   a) D3
   b) Datawrapper
   c) Khartis
   d) **PowerPoint**

13. Which one of these softwares is NOT used to create data visualisation from a dataset?
   a) Rawgraphs
   b) Tableau
   c) **InDesign**
   d) Illustrator

14. In which format export a data visualisation created on an online platform to be able to customise it fully on a vector-based software?
   a) jpg
   b) **svg**
   c) png
   d) csv

15. What is the main downside of using a free online platform?
    a) you can't collaborate with other person
    b) **your data becomes publicly available**
    c) you can't save your work
    d) you can only use a limited amount of data

16. What do we call « sensitive data »?
    a) data that concerns a controversial subject
    b) data about under age persons
    c) data collected without consent
    d) **data that you can't share publicly**

17. In order to choose the good visualisation, you need to focus first on :
    a) **the type and relationship in your data**
    b) the size of your dataset
    c) the source of your data
    d) the liability of your data

18. In Flourish, which one of this format is NOT read by the platform to upload your data ?
    a) csv
    b) json
    c) geojson
    d) **png**

19. In Flourish, what is the basic step to visualise data?
    a) **defining the columns in the table to visualise**
    b) defining the type of data to visualise
    c) defining the ID of the data to visualise
    d) defining the numbers of data to visualise

20. In Flourish, can you upload more than one dataset in a project?
    a) No you can't, you have to create a new project
    b) No you can't, you have to erase the previous one before
    c) **Yes you can, but the second dataset must share at least one column with the previous one**
    d) Yes you can, but the second dataset must contain the same number of rows that the previous one

The main stages of a data project, from data exploration to storytelling

21. Data collection deeply affects the results of a data project. A good data collection would better be described as a matter of external validity or of internal validity?
    a) **External validity**
    b) Internal validity

22. A data analysis project is made of different stages. Which is the last one?
Communicate

23. It is well known that some stages of every data analysis project take much more time than the others. Which ones are they?
Import data and tidy them up

24. Data science is often thought of as a combination of three skills. Which are these three skills?
Statistical knowledge, coding skills, substantive expertise in the concerned domain

25. Which combination of these three skills has been called the “danger zone”?
    a) **Coding skills together with substantive expertise, but without statistical knowledge**
    b) Coding skills together with statistical knowledge, but without substantive experience
    c) Statistical knowledge together with substantive experience, but without coding skills

26. Programming languages for data analysis are the most powerful software available. Because of this reason, they are expensive and can only be afforded by teams with considerable economic resources.
    a) True
    b) **False**

27. Which are the two most commonly used programming languages for data analysis?
    a) Java & R
    b) Python & Ruby
    c) Ruby & Java
    d) **R & Python**

28. Which is one of the main disadvantages of using point-and-click software for data analysis projects?
    a) It takes a lot of time and motivation to start learning it
    b) The graphics produced out-of-the-box are often frustrating
    c) **Each software is specialised in a limited scope of functions**
    d) You need to figure out every little detail of the functions you are using

29. For which reason journalism has a hard time reporting about science?
Because journalists produce well-formed stories with a clear beginning and a clear end, in which the good and the bad can be easily inferred by readers. Science instead describes phenomena in tentative terms, pointing to the limitations of its claims, and refraining from offering value judgements as to what would be better and should be done.)

30. Expressions like “There is no evidence supporting this claim” are ambiguous and can be understood in at least two contrasting ways. Please describe in a few words such ambiguity and produce two examples, one for each contrasting interpretation.
“No evidence” may refer to the temporary lack of confirmation for a reasonable hypothesis, for example about some side-effects of a drug, or to the status of a useless fantasy, such as a conspiracy theory.)

6.3. Module 3 "Tracking and debunking misinformation”

Media ethics

1. When you discover a mistake in the published material
   a) **if you can correct it (eg on the website), do so - and make a note of the changes**
   b) if you can correct it, do so – but do it discreetly
   c) even if you can correct it, you shouldn’t – what’s published is published!
   d) create a new material, with information correcting the mistakes of the previous one

2. When working with data visualisation, take into account:
   a) A technical correctness
   b) Avoiding of potentially misleading composition techniques (eg axis manipulations, aspect ratio, data selection)
   c) **Both of the above**
   d) None of the above

3. “Potter Box” method helps making ethical decisions by:
   a) funny references to the works of J.K. Rowling
   b) consideration of pros and cons
   c) **analysis using four dimensions of dilemma – facts, values, principles and loyalties**
   d) analysis using four dimensions of dilemma – legal, economic, social and cultural

4. Choose a false Big Data statement:
   a) Big Data claims to objectivity and accuracy are misleading
   b) Big Data is always objective knowledge
   c) Big Data is not always better data
   d) not all data are equivalent

Tracking & debunking misinformation

5. What is the Dunning–Kruger effect?
   a) **When people with low knowledge of a topic overestimate their understanding of it**
   b) When by mentioning a topic someone trying to deflect attention ends up attracting more
   c) When someone only believes what confirms their opinion and discredits what doesn’t
   d) When someone is overly influenced by the first information they received on a topic

6. I am interested in a viral post I see on social media but I am not sure if the photo in it is real. I…
   a) Share it and hope someone will tell me
   b) Ask for a source in the comments
   c) Run a reverse search to see if the photo was posted by other people
   d) **Run a reverse search to find the author and context of the photo**

7. Which of these is a recognized global international network of fact-checking organisations?
   a) IFCN
   b) Bellingcat
   c) Reporters sans frontières
   d) Ruptly

8. What is the difference between misinformation, disinformation, and fake news?
   a) **Misinformation is false or misleading information, disinformation is misinformation shared on purpose and fake news are fabricated information given the appearance of news**
   b) Misinformation is unintentional, disinformation is intentional and they are both fake news
   c) Misinformation is slightly false information, disinformation is completely false, and fake news are fabricated information
   d) They are all the same

9. Which of these search engines would be the most appropriate to search a picture allegedly filmed in Russia?
   a) **Yandex**
   b) Google
   c) Baidu
   d) Tineye

10. Where can you get results in chronological order for an image reverse search?
   a) Yandex
   b) Google
   c) Baidu
   d) **Tineye**

11. What is the term for “too many true, false and misleading information circulating during a disease outbreak”?
   a) **Infodemic**
   b) Epidinfo
   c) Pandinfo
   d) Infostorm

12. What is the “Continued-influence effect”?
   a) **When someone keeps believing things even after they’ve been corrected**
   b) When someone cannot stop listening to the same source
   c) When someone cannot get out of their own information bubble
   d) When someone believes something because a lot of people believe it

13. What is Google Lens?
   a) A text recognition and translation tool
   b) A tool to download videos
   c) **A mobile and browser application to perform Image reverse search**
   d) An archiving tool

14. What are the questions you need to answer when researching suspicious images?
    a) What sort of content is it, who posted it on social media, when, where and why was it posted
    b) **What is the content and is it original, who is the author of the content, when was it taken/made, where was it taken and why was the content made and published**
    c) What sort of content is it, when was it posted, why is it important, who is the author and where was it taken
    d) What can you see on the images, where and when were they taken, who posted them on social media and why

15. What is Baidu useful for?
    a) Identifying and translating text from a video in Chinese
    b) Downloading a video from Weibo
    c) Archiving links from Chinese governmental websites
    d) **Running a reverse image search on images seemingly coming from China**

16. Which of these is NOT a reason we believe in false information?
    a) **Low level of education**
    b) It confirms our opinion
    c) It explains a complicated or random event
    d) It comes from an authoritative source

17. What tool can you use to expose a deep fake video?
    a) CrowdTangle
    b) Suncalc
    c) Deepware Scanner
    d) Keyframes in InVID WeVerify

18. What is the best place to check for street level photographs when regular Street view is not available?
    a) Satellites.pro
    b) Deepware Scanner
    c) InVID
    d) Mapillary

19. What is one social network with the least amount of content moderation?
    a) Instagram
    b) Twitter
    c) Telegram
    d) Facebook

20. What is the first step when investigating a video?
    a) Contacting the user who posted it
    b) Doing an image reverse search on screenshots with Invid
    c) Watching it carefully with sound for visual and audio clues
    d) Downloading it in case it is later deleted and archiving the link

21. What are metadata?
    a) **Information about the making of a an image or video such as when, where and on which device it was created**
    b) Proof that an image was manipulated
    c) Edit history of a video
    d) Information about the owner of a website

22. Where can you get metadata from?
    a) A Facebook video
    b) A Whatsapp video
    c) A Twitter video
    d) **The original video sent by email or transfer**

23. Which of these is NOT a reason to archive
    a) To prevent link rot
    b) **To obtain metadata on a photo or video**
    c) To keep reliable traces of posts on social media
    d) To keep track of changes made on websites or articles

24. What tool lets you easily split the video up in order to perform reverse video search?
    a) Google Earth
    b) InVID
    c) BuzzSumo
    d) Deepware Scanner

Social impact of journalism and media transitions

25. What has the internet changed for journalism? Find a statement that is not true.
    a) New intermediaries appeared
    b) New business models appeared
    c) The definition of journalism has blurred
    d) **Information flow regulations have been strengthened**

26. The new logic of internet companies is changing the role of the media - alongside communication tools, making them
    a) **means of data extraction**
    b) means of self-expression
    c) marketing tools
    d) manipulation tools

27. Calculated publics, the term coined by Tarleton Gillespie, refers to dangerous process, in which
    a) **self-perception of the public is based on its algorithmic presentation**
    b) the public only absorbs information that contains figures
    c) the public sphere produces data
    d) people only talk to avatars

28. Algorithms are objective, because…
    a) they are based on mathematical calculations
    b) they are technical creations
    c) they are intelligent
    d) **It is a myth - they are created by people who set specific goals for them, so they are not objective at all!**

29. According to the extractivist paradigm, data mining is like mining coal, because
    a) it is only profitable if you disregard the social costs
    b) it is the basic tool of progress today
    c) building and maintaining its infrastructure is destroying the planet
    d) **a and b**

30. What has the Cambridge Analytica scandal revealed? Please point WRONG answer:
    a) that mass personalization of the manipulated message is possible
    b) that the new media system gives politicians new tools for manipulation
    c) that manipulation practices based on illegally obtained data also took place outside the US
    d) **that cyber corporations respond to such violations on their own initiative**

7. Conclusion

7.1. Video Recording Plan and Initial Guidelines for the Publication of the Training Materials

The initial plan for the programme was to record each session and use recordings as materials for the online part of the course. Throughout the week we used equipment provided by Storytek - a high-quality recording device with both video and sound recording - as well as Zoom recording and streaming that was used by the consortium partners who were not present in the classroom, but wanted to keep track of what was going on in the room. The initial plan was to use these recordings after analysis to create the online course but we realised that the format and quality of these recording were not appropriate for an online learning session (teachers talking to students and not to camera, difficulty to follow the exercises, audio, Q&A from the room not appropriate to online setting, etc.). Lessons were learned from this experience to review lesson recording in the future and review the online course format.

Some other challenges that we spotted were as follows:

- The teachers paid more attention to the students, rather than to the recording device, hence quality of some materials is poor, with teachers being within or outside of the camera's reach;
- Necessity for someone to be in the room at all times to make sure all teachers were aware of the recording, as well as all necessary steps required to record the session were taken.

Based on these observations, as well as a few discussions we had, we came to a conclusion that:

- Recording of the lectures should happen outside of the classroom - we should ask teachers to create a 1:1 version of their course identical or very similar to the one presented in the classroom, based on the materials used there, either during next iterations of the school or in between the second and third training weeks (June/July 2022 and February 2023);
- We need to supplement the material with online-specific content that will adhere to a different way of presenting teaching materials in a self-paced environment.

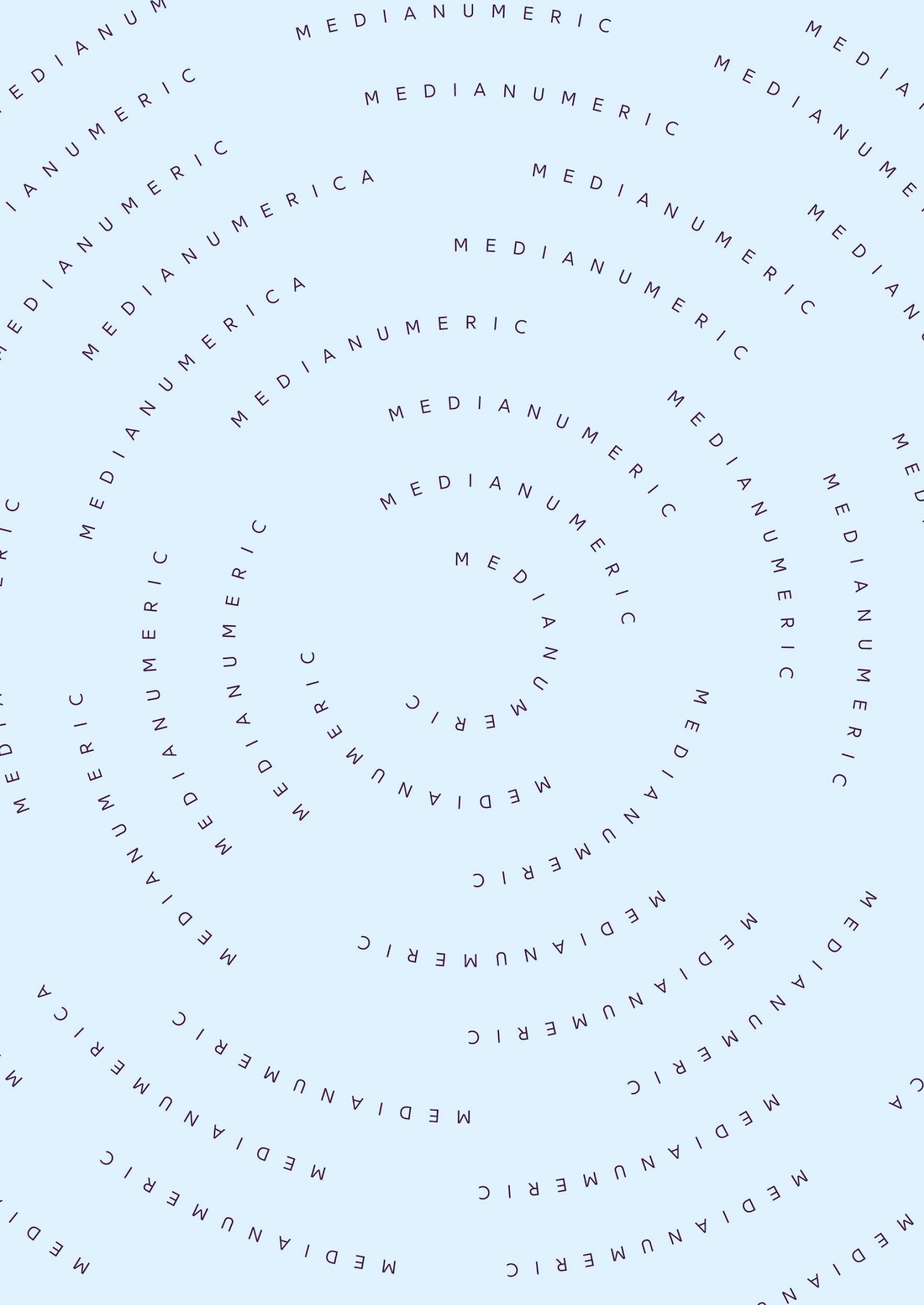
- The recordings for the online format will be finally done in a professional manner during winter school in Warsaw in February 2023.

7.2. Conclusion

Based on several discussions during and after the training programme in The Hague, the group created a *post mortem* document, outlining thoughts and remarks based on the experience gained during the Summer School. Aside from the fact that - for teachers, students and the MediaNumeric partners participating in the event - the entire experience was a high velocity and intense week of learning, we came to the following conclusions:

- While preparing and delivering the programme, we should keep the end goal of the online training course in mind, meaning that eventually some materials will not be fit to use in the online course; that means proper preparation of additional materials, briefing the teachers and potential additional costs;
- More emphasis should be put on workshops and group work;
- Throughout the course there should be more smaller hands-on tasks and more interaction with less lectures;
- Students should have more time to work on the case study;
- The students should work on Google Presentations or Microsoft PowerPoint when preparing their pitches.
- There should be more time to implement and practise the tools shown during the classes;
- Proper technical preparation of the classes is crucial - a technical person should be available at all times to support students' work;
- We should make sure the teachers are aware of the time they're given and that students will probably be asking quite a few questions (so they need to think about their lectures to include time for that);
- We should work on a better structure of providing all lecture materials and handouts to the students via an accessible Google Drive before the classes start;
- We should make sure the teachers know they will have time to show only one tool, the rest should be mentioned in the handouts (links to additional tools, links to the tutorials, etc.);
- We should make sure that the students know what's expected of them regarding the case study - not only mention that during Day 0, but also repeat that during the first day of the school and throughout the week;
- Lectures on Media Ethics and Social Impact of Journalism should take the form of a keynote lecture kicking off the course;
- The students should have some time off, ie. on Wednesday: finish around 2-3pm and give them time to do whatever they'd like (ie. network, work);
- The programme should allocate lots of time for practical exercises in the afternoon;
- The programme should allocate time and space for revision of the materials prior to the quizzes;
- The quizzes covering past classes should take place at least one day after the last lecture that the test will cover to give the students time to revise the materials;

- We should have each case study session start with a small task/exercise;
- We should make sure that all materials are kept consistent and that the teachers use templates provided by the MediaNumeric team. This practice was implemented for The Hague summer school and it made everybody's work much easier;
- We should make sure that all students have access to, and have explored, the Google Drive before classes start;
- Direct live communication with students (Basecamp, Slack or WhatsSpp group) - to be considered.


MEDIANUMERIC