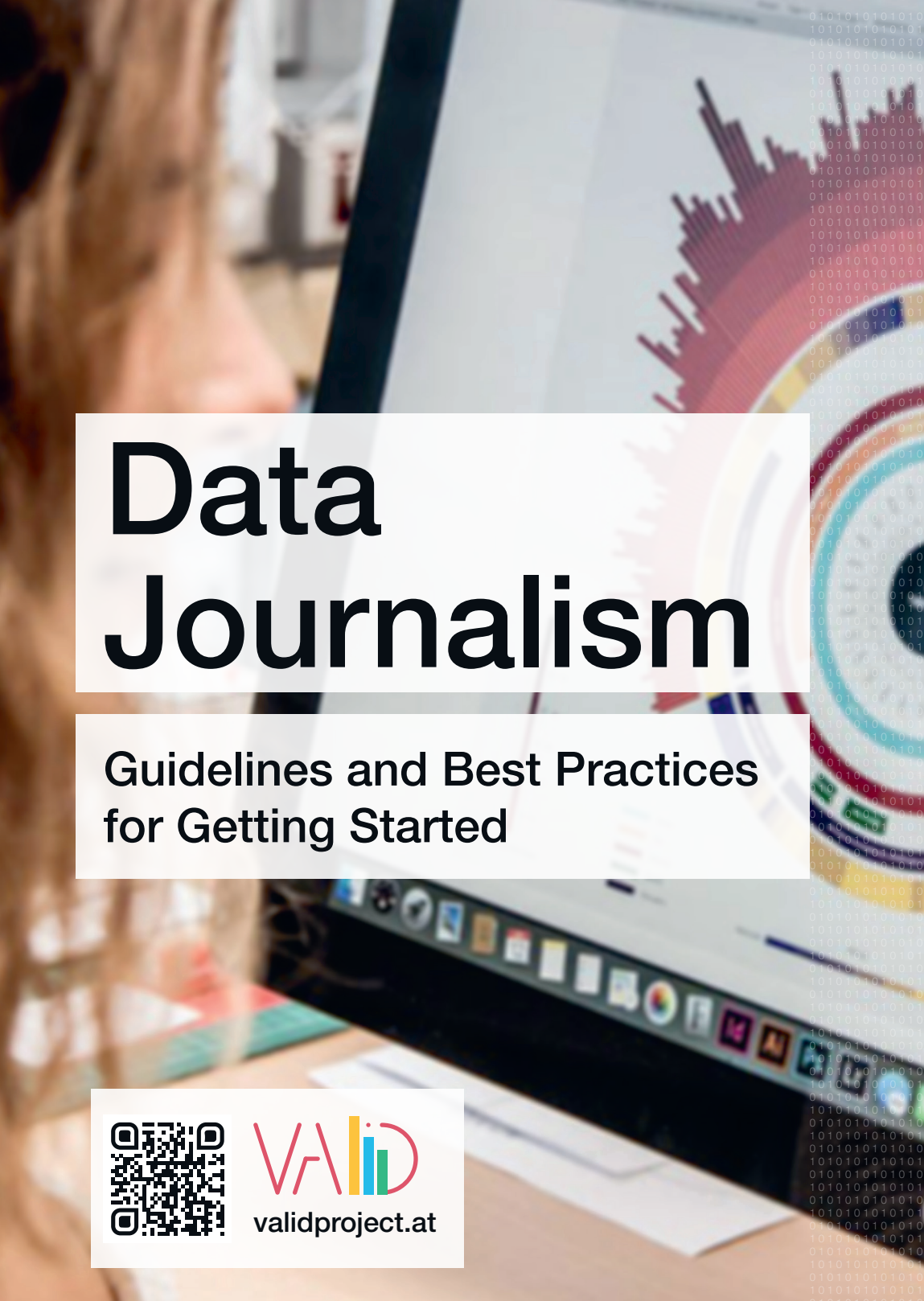

# Data Journalism

## Guidelines and Best Practices for Getting Started

validproject.at

Deliverable: D1.3

Contributors:
Wolfgang Aigner | Eva Goldgruber | Florian Grassinger | Robert Gutounig | Alexander Rind
Michael Sedlmair | Christina Stoiber

Titelfoto: FH JOANNEUM / Manfred Terler

November 2018

Project title:
Visual Analytics in Data-driven Journalism
http://www.validproject.at/

This work was supported by the Austrian Ministry for Transport, Innovation and Technology (BMVIT) under the ICT of the future program via the VALiD project no. 845598.



This booklet was created in the context of the research project “Visual Analytics in Data-Driven Journalism (VALiD)”:

Project Synopsis:
With the ever-growing amount and availability of data, it becomes crucial for journalists to use elements of data science in their work. This trend led to the advent of the emerging field of Data-Driven Journalism (DDJ), which involves computer-supported data-based reasoning as well as interactive visualization.

The goal of the research project „VALiD - Visual Analytics in Data-Driven Journalism“ is to design methods and tools to support data journalists in dealing with complex heterogeneous data and identify best practices of how to integrate DDJ into newsrooms. The developed approaches can be used, for example, to analyze data from parliamentary speeches or money flows between organizations.

# Contents

# Preface

This document represents a practical summary from experiences and results gained during the research project “VALiD - Visual Analytics in Data-Driven Journalism”. It addresses some of the major issues around data journalism practice. A large part is taken from a review of the research literature review on this topic. You can find the references for this review and links to other project-related resources at the end of this document. This document addresses journalists and communicators who have little to no experience in the field of data-intensive newswork.

# Why data journalism?

Data journalism is the systematic collection, analysis and visualization of structured information using algorithms and quantitative social science methods in the making and presentation of news stories. Many see this phenomenon as journalism’s response to an increasingly data-dependent society. Data journalism is also seen by many media experts as one of the key elements of the future of journalism. Compared to ‘ordinary’ text journalism, it provides additional value for recipients. The reader’s time spent on data-intensive projects is usually longer and data stories are shared more frequently with others. However, data-intensive newswork is still rare as a daily practice in newsrooms, especially in smaller and medium-sized media organizations.

# Where to find inspiring examples of data journalism?

Prototypical examples of data journalism include the Guardian's ‘Afghan’[1] and ‘Iraq War Logs’[2] or the ‘Toxic Waters project’[3] from the New York Times, which examined pollution in American waters. Along more artistic lines there is the Wall Street Journal´s rhyme analysis[4] of the Broadway musical ‘Hamilton’.

**LINK TIP:**
**If you want to look for the latest best practice examples regarding data journalism check out the nominees for the Data Journalism Awards.**
https://www.datajournalismawards.org

## What does successful collaboration in data journalism look like?

Data journalism is often a much more collaborative effort than traditional journalism. Numerous experts help to produce the final piece, as usually there is no "data journalist" working on his or her own. In many cases, a team of at least three experts has proven most promising in practice: a programmer, a designer and an author. Complementary backgrounds are key for such an interdisciplinary multimedia project. Open collaboration between these disciplines has turned out to make the end product easier to use and more understandable for the reader.
Cross-disciplinary collaboration and mutual support are essential in data journalism and happen also between different media outlets. So whenever data skills are coming to an end, one should try to find out who might have these skills or try to connect with other data journalists (e.g. at meetups[5] or data journalism communities[6]). Also, technical problems or questions concerning content are shared on platforms such as the messaging service Twitter. Twitter is a central dissemination and feedback channel where the hashtag #ddj is used to connect people interested in the subject. For technical issues, the Q&A site Stack Overflow[7] might also be the place to go and ask your question or search for similar questions that might already have been answered.

## What are suitable topics for data journalism projects?

Data journalistic projects are carried out in a broad variety of topics, including (but not restricted to):

- **Demographics**
  Example: New Berliners and native Berliners by Berliner Morgenpost
  https://www.morgenpost.de/article136530429

[1] https://www.theguardian.com/world/the-war-logs
[2] https://www.theguardian.com/world/iraq-war-logs
[3] https://www.nytimes.com/interactive/projects/toxic-waters
[4] http://graphics.wsj.com/hamilton/
[5] A good place to start may be the local meetups of the Hacks/Hackers movement: https://hackshackers.com/ or look out for local initiatives such as http://ddj-monaco.de/.
[6] Data Journalism Den (https://datajournalismden.org) provides both exchange as well as matchmaking for potential projects.
[7] https://stackoverflow.com/

- **Crime**
  Example: Crime in Context by The Marshall Project
  https://www.themarshallproject.org/2016/08/18/crime-in-context

- **Politics**
  Example: Electionland by ProPublica and The Electionland Coalition
  https://www.propublica.org/electionland

- **Business/economy/finances**
  Example: The Panama Papers by International Consortium of Investigative Journalists | https://www.icij.org/investigations/panama-papers/

- **Healthcare**
  Example: Death & Dysfunction by NJ.com | http://death.nj.com/

- **School/education**
  Example: Education data to empower parents, pupils, educators & build accountability by Passmark | http://passmark.org.za/

- **Environment**
  Example: How Much Hotter Is Your Hometown Than When You Were Born? by The New York Times | https://nyti.ms/2MllbH5

- **Job/working conditions**
  Example: The Jobless Rate for People Like You by The New York Times
  https://archive.nytimes.com/www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html

- **Urban services and transportation**
  Example: Bike Sharing Atlas by the University of Vienna
  http://bikesharingatlas.org

- **Culture**
  The Rhymes behind Hamilton by The Wall Street Journal
  http://graphics.wsj.com/hamilton/

- **Sports**
  Example: LeBron James Scores 5,995th Playoff Point, Taking the Record From Michael Jordan by The New York Times
  https://nyti.ms/2s1yRz2

## What are the data journalism workflows?

There are two primary ways to start a data journalism story: (1) a dataset serves as the initial starting point for the story, (2) a dataset provides more information for a story topic that was already present. In case (2) the process for data journalists can be described as follows (Aitamurto, T., Sirkkunen, E. & Lehtonen, P. 2011):

1. Identify the gist for the story and the potential role for the data in the story.

2. Identify and obtain the right datasets to respond to journalists' questions (see chapter „Getting the Data").

3. Modify the data to make them ready for analysis — e.g., correcting errors in the datasets (see chapter „Preparing and Cleaning the Data").

4. Analyzing the data with the right tools and mashing the data with other datasets if relevant (see chapter „Analyzing the Data").

5. Produce the story: text, visualizations, interactive elements (see chapter „Build the Data Story and Visualize the Data").

6. Publish the datasets that were used in the analysis (see chapter „Publish the Product").

7. Invite readers to participate by reusing the data, commenting on and sharing the story through applications in social media and submitting more content through applications such as Facebook or Twitter.

## Getting the Data

Data can, for instance, be obtained from the following sources:

- **Open government data portals / pan-national organizations**
  Many governments on national or regional levels as well as pan-national organizations (such as the European Union or the various agencies of the United Nations) run their own Open Government Data portals or initiatives.

**LINK TIPS:**
European Data Portal: https://www.europeandataportal.eu
OECD: http://stats.oecd.org/
World Bank: http://data.worldbank.org/
UNHCR: http://popstats.unhcr.org/en/overview
WHO: http://www.who.int/gho/en/
Open Government Data for Austria: http://data.gv.at

- **Freedom of Information Request / Access to Information Request**
  Freedom of Information (FOI) is a fundamental right to public access to documents and files of public administration. Within this framework, for example, offices and authorities may be obliged to publish their files and procedures (principle of public access) or to make them accessible (administrative transparency) and to define binding quality standards for access for this purpose.

**LINK TIPS:**
There are countries which issued Freedom of Information Acts, e.g.
- USA: https://www.foia.gov/
- Great Britain: https://ico.org.uk/
- Germany: https://www.bfdi.bund.de

- **Journalist's own research** (compiling surveys, polls etc.)
  In case no suitable dataset is at hand journalists can use their own network or even their media organization's audience to collect data by using a survey. In this case, the audience's interest in the results can also be presupposed.

- **Crowdsourcing: Collecting data from the audience or engaging them in a topic**
  The potential audience itself can offer valuable support in gathering data through crowdsourcing activities. The crowdsourcing campaign is at the same time a pre-test to see if there is enough interest in a particular subject and can help to build trust with the readers.

**EXAMPLE:**
Free the files: https://www.propublica.org/series/free-the-files

- **Leaks**
  Leaks of unpublished documents by both governmental and non-governmental organizations have become a major resource for data journalistic projects.

**EXAMPLES:**
Iraq – the war logs: https://www.theguardian.com/world/iraq-war-logs
Panama Papers: https://www.icij.org/investigations/panama-papers/

- **Non-commercial organizations: universities, research institutions, NGOs etc**.
  Numerous organizations publish a variety of data sources which can be used for journalistic analysis. Journalists can also draw from data compiled for academic research provided through data archives.

**LINK TIP:**
The Austrian Social Science Data Archive: https://aussda.at

- **Companies**
  Data from companies include annual reports, market research etc. Data journalists sometimes voice scepticism concerning the reliability of these datasets. It is advisable to check the circumstances and the methods of data collection before completely relying on it.

## Preparing and Cleaning the Data

In order to be able to explore, analyze, or visualize data, it needs to be available in a structured, machine-readable format (e.g. CSV, XLS, JSON, XML, plain text etc.). However, even where access to data is in principle possible, it is often not given in forms and formats that allow for an easy processing with computer tools. A prototypical example of this is the publication of data in form of PDF reports, where it is very difficult to extract the data in a machine-readable format. In these cases, technical methods such as data scraping or data mining need to be applied.

**PRACTICAL TIP:**
Extracting Data from PDFs:
http://okfnlabs.org/blog/2016/04/19/pdf-tools-extract-text-and-data-from-pdfs.html

Having obtained the right datasets they have to be filtered and/or segments selected which are relevant for the story. Furthermore, datasets almost always include errors. Journalists have to clean them up before drawing conclusions.

## Analyzing the Data

An essential element when analyzing complex datasets is the use of Visual Analytics (VA) in order to identify patterns or outliers with the goal of finding a story within the data. Analyzing data may range from simply comparing values (e.g. the number of men vs. the number of women) to changes over time (e.g. temperature over time) to sophisticated statistical methods, data mining etc. An often applied method, especially for investigating social-media data, is social-network analysis. Some methodologies even allow the analysis of large amounts of unstructured data such as text (e.g. sentiment analysis).

**See tools for data analysis in the Data Journalism Toolbox at the end of this document.**

# Build the Data Story and Visualize the Data

A key element of data journalism is the visualization of the data. Numerous helpful and free tools exist that can help journalists produce appealing graphics and visualizations by themselves without the help of professionals such as designers or programmers. There are lots of different ways to present your data. Most commonly, data stories are visualized with one or more of the following:

- (Static) pictures
- (Simple static) charts
- Maps (static or interactive)
- Tables
- Networks
- Infographics
- Timelines
- Animated visualisations
- Lists
- Sound and video
- ...

When designing an interactive infographic several factors have to be taken into account. Psychological research on visual salience found that exceptions from the norm attract the viewer's attention and can be used as a stylistic device. It is also recommended to start the visualization process with a storyboard, which illustrates the graphic as a sequence of arranged images. This helps journalists and developers in their communication with each other as well as the production process. Other useful resources for projects are wireframes, sketches and drafts drawn on a computer. Those shall help to better integrate the view of the end user into the concept of the visualization.

An important step in building a visualization for a story is how to balance open exploration with a "prescriptive narrative approach". Although data journalism is often associated with exploratory visualizations, guided narration still seems to be preferred by the majority of readers. Such guidance can be provided through animated transitions between pre-defined views with text annotations highlighting the relevant insights. These data stories can progress as the reader scrolls down, clicks on a button or even automatically such as in a video. Often the individual views of such a story allow the reader to explore further for more details.

**EXAMPLE:**
New York Times, LeBron Career Playoff Point Record: https://nyti.ms/2s1yRz2

**See tools for data vizualizations in the Data Journalism Toolbox at the end of this document.**

## Revising and Editing

Before visualizations are published, (internal) user tests should be performed to ensure a high level of usability as well as other feedback loops, also taking into account the question of accessibility, e.g. for visually impaired users. A further important step is the adaption of visualizations for different platforms, operating systems, browsers and also mobile use.

## Publish the Product

Data stories need to be highlighted to get the attention they deserve. A prominent display on the welcome page of the medium as well as sharing it via social media is important. In addition, the placement on the website should not be underestimated. A prominent setup is important as pageviews, retention time and the number of viewers are central metrics for measuring the success of data-driven projects.

## Open Data

Online journalism, in general, facilitates interactivity and engagement. In data journalism, after publishing the projects, the data teams often publish the raw datasets or the source code of their programs (e.g. on GitHub[8]). In case the dataset you use for your analysis is not yet public, it is recommended that you publish it if no legal reasons speak against it. This also contributes to the transparency of your work and increases trust among your audience.

[8] https://github.com/

## Where to pick up data journalism skills?

If you want to pick up data journalism skills look for training in one (or more) of these areas:

- Digital journalism
- Structured journalism & knowledge representation
- Web programming (e.g. JavaScript)/development for the cloud
- Data visualization/infographics
- Data analysis
- Data mining
- Text mining (e.g. NLP)
- Network analysis
- Web design

As a lot of data journalistic projects are development projects, skills for (agile) project management are also highly recommended. However, journalistic skills are still the vital key element in data journalism and serve as a basis for all data stories.

**LINK TIPS:**

- National Institute for Computer Assisted Reporting (NICAR) offers a range of training activities from workshops, conferences to online training: https://ire.org
- A free online course on data journalism can be found on the learning platform https://learno.net/courses/doing-journalism-with-data-first-steps-skills-and-tools
- The research center CORRECTIV offers an online course for local journalists (In German): https://correctiv.org/bildung/ddj/datenjournalismus-fuer-lokalreporter/einfuehrung-was-ist-datenjournalismus/
- The association DOSSIER: Academy is offering training on data journalism skills (in German). Check out the next dates at https://academy.dossier.at/

# Data Journalism Toolbox

## Data Sources

- The Austrian Social Science Data Archive: https://aussda.at
- European Data Portal: https://www.europeandataportal.eu
- OECD: http://stats.oecd.org/
- Open Government Data for Austria: http://data.gv.at
- UNHCR: http://popstats.unhcr.org/en/overview
- WHO: http://www.who.int/gho/en/
- WikiLeaks: https://wikileaks.org
- World Bank: http://data.worldbank.org/

## Tool Overviews

- Data Journalism Tools: https://datajournalism.tools/
- Data Libraries: https://infovis-wiki.net/wiki/Data_Libraries

## Tools for Data Analysis

- Microsoft Excel/LibreOffice Calc/Google Spreadsheet
- Netflower: http://netflower.fhstp.ac.at/
- OpenRefine: http://openrefine.org/
- Breve: http://hdlab.stanford.edu/breve/
- Trifacta Wrangler https://www.trifacta.com/start-wrangling/
- Orange: https://orange.biolab.si/
- Keshif: https://keshif.me/

## Tools for Preparing Data

- Extracting Data from PDFs: http://okfnlabs.org/blog/2016/04/19/pdf-tools-extract-text-and-data-from-pdfs.html
- Converting popular data formats: https://www.csvjson.com/

## Tools for Visualizations

- Caspio: https://www.caspio.com/
- Tableau Software: https://www.tableau.com
- Datawrapper: https://www.datawrapper.de/
- RAW Graphics: https://rawgraphs.io/
- Google Fusion Tables: google.com/fusiontables
- Carto: https://carto.com/
- Highcharts: https://www.highcharts.com

## Training

- CORRECTIV (in German): https://correctiv.org/bildung/ddj/datenjournalismus-fuer-lokalreporter/einfuehrung-was-ist-datenjournalismus/
- DOSSIER: Academy (in German): https://academy.dossier.at/
- LEARNO.NET: https://learno.net/courses/doing-journalism-with-data-first-steps-skills-and-tools
- LinkedIn Learning: https://www.linkedin.com/learning
- National Institute for Computer Assisted Reporting (NICAR): https://ire.org
- SAGE Campus: https://campus.sagepub.com

**READING TIP:**
A good place for next steps on data journalism is the Data Journalism Handbook which can be found online at https://datajournalismhandbook.org/

# Glossary

**Agile Development**
Agile development refers to approaches which should increase transparency and flexibility and lead to a faster use of the developed systems in order to minimize risks in the development process. The goal of agile development is to make the development process more flexible and streamlined than is the case with classic process models. An example of such a method is 'Scrum'[9].

**Algorithm**
An algorithm is a step-by-step recipe for solving a problem or a class of problems. Algorithms consist of many well-defined single steps.

**Crowdsourcing**
Crowdsourcing refers to the outsourcing of tasks to a group of users, usually over the Internet. It splits the work among the participants in order to obtain a cumulative result.

**CSS**
Cascading Style Sheets, or CSS for short, is a stylesheet language for electronic documents and, together with HTML, one of the main languages of the World Wide Web.

**CSV**
The CSV file format stands for comma-separated values and describes the structure of a text file for storing or exchanging simply structured data. In CSV files, tables or a list of lists of different lengths can be displayed.

**Data scraping**
Data scraping is a computer technology in which data is extracted from an online document and used as input into another program or stored in a structured data file.

**Data Story**
Data stories are stories that are data driven. They start from a narrative based on or containing a narrative connected to data. They typically include evidence for the data in the form of graphics or (interactive) data visualizations.

[9] https://www.scrum.org

**Freedom of Information Request**
Freedom of Information Acts are legal institutions that enable citizens to know what administrative information is held by the government and that the government is obliged to disclose it. This principle has been elaborated in the legislation of some countries (e.g. USA, Great Britain & Australia).

**HTML**
The Hypertext Markup Language, abbreviated HTML, is a text-based markup language for structuring digital documents such as texts with hyperlinks, images and other content. HTML documents are the basis of the World Wide Web and are displayed by web browsers.

**Infographics**
Infographics are used for visual communication of concepts, processes or data in an image. In addition to the two classical disciplines of text and photo journalism, it is an independent form of journalistic presentation that visually presents information.

**JavaScript**
JavaScript (JS) is a script language for evaluating user interactions, modifying, reloading or generating content and thus expands the possibilities of HTML and CSS. JavaScript can be embedded into webpage and executed by browsers.

**JSON**
The JavaScript Object Notation, short JSON, is a compact data format in an easily readable text form for the purpose of data exchange between applications.

**NLP**
Natural Language Processing: computational linguistics examines how natural language can be algorithmically processed with the help of computers in the form of text or speech data. It is the interface between linguistics and computer science.

**Open Data, Open Government Data**
Open data is data that may be used, disseminated and reused by anyone without any restrictions. If such data is provided by a government body it is also called ‘Open Government Data’.

**Pageviews, page impressions**
Page impressions refer to the number of times a single web page is viewed using a web browser.

**Retention time**
Retention time is the amount of time a visitor stays on a particular web page or an entire website. The purpose of web analytics is to provide the website operator with information on how intensively users are interested in the website's content and how user-friendly it is.

**Social network analysis**
Social network analysis is an empirical social research method for recording and analyzing social relationships and social networks. Social network analysis propagates a view of social phenomena that emphasizes their relational character. Connections and interdependencies between units (e.g. persons or organizations) are in the foreground, not their individual attributes and properties.

**SQL**
SQL is a database language for defining data structures in relational databases and for editing (inserting, modifying, deleting) and querying databases based on them.

**Visualization**
Visualization refers to the process of transforming data (e.g. numbers or texts) into a graphic or visually comprehensible form.

**Visual Analytics**
Visual Analytics is an interdisciplinary approach that combines the advantages of different research areas. The goal of the Visual Analytics method is to gain insights from extremely large and complex datasets. The approach combines the strengths of automatic data analysis with a human's ability to quickly visualize patterns or trends. Through appropriate interaction mechanisms, data can be visually explored and insights gained.

**Wireframe**
The term wireframe is used to describe a very early conceptual design of a website or software front end. The design and function do not yet play a role. The focus is on the arrangement of elements and user guidance (UX, user experience).

## Resources

These guidelines are based on the results of the research project VALID - Visual Analytics in Data-Driven Journalism. You can find the related documents here: http://www.validproject.at

## Bibliography


- Aitamurto, T., Sirkkunen, E., & Lehtonen, P. (2011). Trends in data journalism. http://virtual.vtt.fi/virtual/nextmedia/Deliverables-2011/D3.2.1.2.B_Hyperlocal_Trends_In%20Data_Journalism.pdf

- Ausserhofer, J., Gutounig, R., Oppermann, M., Matiasek, S., & Goldgruber, E. (2017). The datafication of data journalism scholarship: Focal points, methods and research propositions for the investigation of data-intensive newswork. Journalism. https://doi.org/10.1177/1464884917700667

- Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2017). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. Journalism: Theory, Practice & Criticism, 146488491773569. https://doi.org/10.1177/1464884917735691

- Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

validproject.at