**Process of data journalism**

The process to transform raw data into stories is akin to a refinement and transformation. The main goal is to extract information recipients can act upon. The task of a data journalist is to extract what is hidden. This approach can be applied to almost any context, such as finances, health, environment or other areas of public interest.

**Inverted pyramid of data journalism**

In 2011, Paul Bradshaw introduced a model, he called ["The Inverted Pyramid of Data Journalism"](http://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism/).

**Steps of the process**

In order to achieve this, the process should be split up into several steps. While the steps leading to results can differ, a basic distinction can be made by looking at six phases:

1. Find: Searching for data on the web
2. Clean: Process to filter and transform data, preparation for visualization
3. Visualize: Displaying the pattern, either as a static or animated visual
4. Publish: Integrating the visuals, attaching data to stories
5. Distribute: Enabling access on a variety of devices, such as the web, tablets and mobile
6. Measure: Tracking usage of data stories over time and across the spectrum of uses.

**Description of the steps**

**Finding data**

Data can be obtained directly from governmental databases such as [data.gov](https://en.wikipedia.org/wiki/Data.gov), [data.gov.uk](https://en.wikipedia.org/wiki/Data.gov.uk) and World Bank Data APIbut also by placing [Freedom of Information requests](https://en.wikipedia.org/wiki/Freedom_of_Information_request) to government agencies; some requests are made and aggregated on websites like the UK's What Do They Know. While there is a worldwide trend towards opening data, there are national differences as to what extent that information is freely available in usable formats. If the data is in a webpage, scrapers are used to generate a spreadsheet. Examples of scrapers are: Import.io, [ScraperWiki](https://en.wikipedia.org/wiki/ScraperWiki" \o "ScraperWiki), [OutWit Hub](https://en.wikipedia.org/wiki/OutWit_Hub" \o "OutWit Hub) and [Needlebase](https://en.wikipedia.org/w/index.php?title=Needlebase&action=edit&redlink=1) (retired in 2012). In other cases OCR software can be used to get data from PDFs.

Data can also be created by the public through crowd sourcing, as shown in March 2012 at the Datajournalism Conference in Hamburg by Henk van Ess.

**Cleaning data**

Usually data is not in a format that is easy to visualize. Examples are that there are too many data points or that the rows and columns need to be sorted differently. Another issue is that once investigated many datasets need to be cleaned, structured and transformed. Various tools like [Google Refine](https://en.wikipedia.org/wiki/Google_Refine) ([open source](https://en.wikipedia.org/wiki/Open-source_software)), [Data Wrangler](https://en.wikipedia.org/w/index.php?title=Data_Wrangler&action=edit&redlink=1) and [Google Spreadsheets](https://en.wikipedia.org/wiki/Google_Spreadsheet) allow uploading, extracting or formatting data.

**Visualizing data**

To visualize data in the form of graphs and charts, applications such as [Many Eyes](https://en.wikipedia.org/wiki/Many_Eyes) or [Tableau Public](https://en.wikipedia.org/wiki/Tableau_Public) are available. [Yahoo! Pipes](https://en.wikipedia.org/wiki/Yahoo!_Pipes) and Open Heat Map are examples of tools that enable the creation of maps based on data spreadsheets. The number of options and platforms is expanding. Some new offerings provide options to search, display and embed data, an example being [Timetric](https://en.wikipedia.org/w/index.php?title=Timetric&action=edit&redlink=1" \o "Timetric (page does not exist)).

To create meaningful and relevant visualizations, journalists use a growing number of tools. There are by now, several descriptions what to look for and how to do it. Most notable published articles are:

* Joel Gunter: "#ijf11: Lessons in data journalism from the New York Times"[
* Steve Myers: "Using Data Visualization as a Reporting Tool Can Reveal Story’s Shape", including a link to a tutorial by Sarah Cohen

As of 2011, the use of HTML 5 libraries using the [canvas](https://en.wikipedia.org/wiki/Canvas) tag is gaining in popularity. There are numerous libraries enabling to graph data in a growing variety of forms. One example is [RGraph](https://en.wikipedia.org/wiki/RGraph" \o "RGraph).  As of 2011 there is a growing list of JavaScript libraries allowing to visualize data.

**Publishing data story**[

There are different options to publish data and visualizations. A basic approach is to attach the data to single stories, similar to embedding web videos. More advanced concepts allow to create single dossiers, e.g. to display a number of visualizations, articles and links to the data on one page. Often such specials have to be coded individually, as many Content Management Systems are designed to display single posts based on the date of publication.

**Distributing data**

Providing access to existing data is another phase, which is gaining importance. Think of the sites as "marketplaces" (commercial or not), where datasets can be found easily by others. Especially of the insights for an article where gained from Open Data, journalists should provide a link to the data they used for others to investigate (potentially starting another cycle of interrogation, leading to new insights).

Providing access to data and enabling groups to discuss what information could be extracted is the main idea behind Buzzdata, a site using the concepts of social media such as sharing and following to create a community for data investigations.

Other platforms (which can be used both to gather or to distribute data):

* Help Me Investigate (created by Paul Bradshaw)
* Timetric
* ScraperWiki

**Measuring the impact of data stories**

A final step of the process is to measure how often a dataset or visualization is viewed.

In the context of data-driven journalism, the extent of such tracking, such as collecting user data or any other information that could be used for marketing reasons or other uses beyond the control of the user, should be viewed as problematic.One newer, non-intrusive option to measure usage is a lightweight tracker called PixelPing. The tracker is the result of a project by [ProPublica](https://en.wikipedia.org/wiki/ProPublica" \o "ProPublica) and [DocumentCloud](https://en.wikipedia.org/wiki/DocumentCloud" \o "DocumentCloud).There is a corresponding service to collect the data. The software is open source and can be downloaded via GitHub.

Examples

There is a growing list of examples how data-driven journalism can be applied. *The Guardian*, one of the pioneering media companies in this space (see "Data journalism at the Guardian: what is it and how do we do it?"), has compiled an extensive list of data stories, see: "All of our data journalism in one spreadsheet".

Other prominent uses of data-driven journalism are related to the release by whistle-blower organization [WikiLeaks](https://en.wikipedia.org/wiki/WikiLeaks" \o "WikiLeaks) of the [Afghan War Diary](https://en.wikipedia.org/wiki/2010_Afghan_War_documents_leak), a compendium of 91,000 secret military reports covering the war in Afghanistan from 2004 to 2010.  Three global broadsheets, namely [*The Guardian*](https://en.wikipedia.org/wiki/The_Guardian), [*The New York Times*](https://en.wikipedia.org/wiki/The_New_York_Times) and [*Der Spiegel*](https://en.wikipedia.org/wiki/Der_Spiegel), dedicated extensive sectionsto the documents; [The Guardian](https://en.wikipedia.org/wiki/The_Guardian)'s reporting included an interactive map pointing out the type, location and casualties caused by 16,000 [IED](https://en.wikipedia.org/wiki/Improvised_explosive_device) attacks, [The New York Times](https://en.wikipedia.org/wiki/The_New_York_Times) published a selection of reports that permits rolling over underlined text to reveal explanations of military terms,  while [Der Spiegel](https://en.wikipedia.org/wiki/Der_Spiegel) provided hybrid visualizations (containing both graphs and maps) on topics like the number deaths related to insurgent bomb attacks. For the [Iraq War logs release](https://en.wikipedia.org/wiki/Iraq_War_documents_leak), *The Guardian* used [Google Fusion Tables](https://en.wikipedia.org/wiki/Google_Fusion_Tables) to create an interactive map of every incident where someone died, a technique it used again in the [England riots](https://en.wikipedia.org/wiki/2011_England_riots) of 2011.