DATA
JOURNALISM

An Introductory Guidebook

DATA
JOURNALISM

ALJAZEERA
MEDIA INSTITUTE

TABLE
OF CONTENTS

1. INTRODUCTION

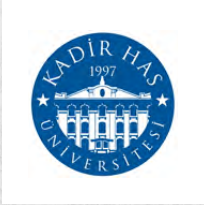Data Journalism is a flourishing field of journalism that actively looks for, analyses and interprets various forms of data for storytelling.

More than half of all news organizations in the US and Europe now have at least one dedicated data journalist working in their newsrooms.
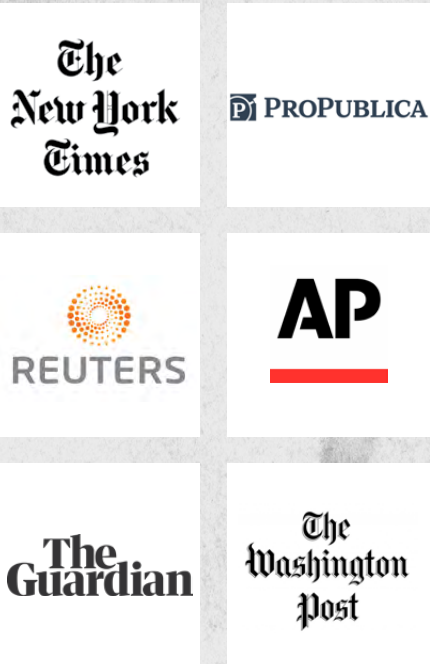
An assortment of newsrooms with dedicated data journalists

There are also now over two hundred data journalism modules and programs being taught in journalism schools all over the world. These programs are set to prepare journalists for the next wave of analytical and accountability-driven journalism.

For the average global citizen, more and more of our everyday lives are impacted by computers and data. In order to hold power to account, journalists should be empowered with the skills and tools to make sense of this data.

By working through real case studies within Al Jazeera, we hope to provide readers with a concise and practical introduction to data journalism as seen through the lens of an Arab audience.

This guidebook will move beyond the technical details and rather focus on helping you think critically and creatively about telling human stories hidden within data.

Whether you're a novice or experienced journalist, this guidebook is for you.

An assortment of universities with data journalism modules and programs.

2. TELLING UNTOLD STORIES

At its core, the principles of data journalism and traditional journalism are the same. Both involve asking questions to paint a truthful and accurate picture of the world.

In the case of traditional media, this information is usually obtained from an individual person's account of a story (a single data point). Data journalism takes this one step further by asking questions to a collection of data points (many people's stories). This approach can help reveal the larger context of a story.

Journalism

Untold stories

Data analysis

Data Journalism is the intersection between journalism and data analysis - often through the use of technology
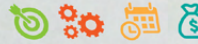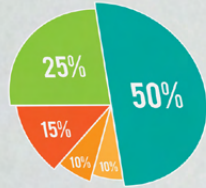
a. What is data journalism?

When people hear the phrase "data journalism" most automatically think of charts and infographics.

However, data journalism is a larger field. It's the entire process of deriving meaning from data to develop a story - not only the visual output.

A written story that relies on data analysis and interpretation is a better example of data journalism than an infographic with dozens of meaningless numbers. The key ingredient is asking questions to our data just as if we were interviewing it.
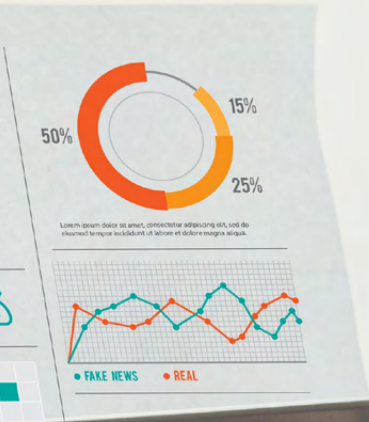
In brief:

Data journalism is a form of journalism where your interview subject is data.

This means that a data journalism story is often **PLATFORM INDEPENDENT**

MAP

VIDEO
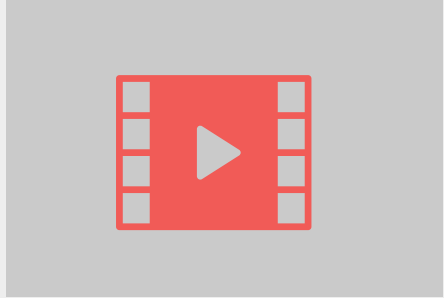
CHART

ARTICLE

Your output can take on the form of a map, video, chart, written article and even social media posts. This allows you to be very creative with your output and not be constrained by a specific medium. This cross-platform approach is a very important part of digital content creation.

Data-led stories have the power of reaching and engaging with new audiences by making sense of the data-rich world that we live in.

It is important to remember that data journalism is not about using shiny new technologies, rather it is about using technology to help extract contextual information for your readers.

b. 5 myths about data journalism

We asked a few journalists from our newsroom to share some of the biggest myths about what data journalists do. Here's what they had to say:

⊘ MYTH	✓ REALITY
"That data journalism isn't personal."	More than anything, data is about stories that play a direct role in people's lives. Many data stories have the ability to not only tell individual stories but also contextualise a story by placing a person in his or her neighbourhood or country.
"Data journalists are not real journalists, they're only interested in numbers, not telling stories."	Data journalists do more than just sift through mountains of data and identify trends. They can often provide a creative perspective to a story which allows them to engage with new audiences.
"Data journalism is for programmers and designers"	While it's true that data-driven stories can benefit from people with technical and design skills, most of the work stems from an editorial understanding of a subject. As long as you have an eye for a story and are willing to collaborate with others, you can become a good data journalist.
"Data journalism is all about making charts and infographics."	An infographic or chart without an underlying story is not data journalism. A data journalism project should involve the uncovering of a story from a dataset.
"Data journalism is expensive and time-consuming."	Within Al Jazeera we've produced award-winning data journalism projects with a mobile phone, camera and computer (as will be demonstrated in our case study below). While, longer-term investigative projects may take time and resources to develop, there are many daily stories that newsrooms are producing that involve analysing and presenting data.

C. Searching for facts

The spread of misinformation online has created a huge problem for news consumers. Building your audience's trust in your data stories requires that you treat your data sources like every other source of information.

Your data must be verified for accuracy and truth.

Data ≠ Truth

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Tips for sourcing reliable data:

1. Use trustworthy sources

Finding good data sources can be difficult. However, it's always better to build upon an incomplete data source that you can trust rather than blindly using a complete dataset that you cannot.

2. Cross-reference datasets

Always cross-reference your data with news stories, reports or other data. This will save you a lot of time and effort once it comes to cleaning and analysing your data.

3. Watch out for missing or outdated values

Just as information becomes outdated so too does data. Always try to find the most relevant and up-to-date dataset to ensure that your story best represents the current facts.

4. Understand the data collection methodology

People are often responsible for gathering raw data. To avoid unnecessary errors always ask yourself how was this data collected and what does it actually represent?

5. What are the consequences of getting it wrong?

Data-driven stories have a big advantage over traditional stories in that they can more objectively be validated and confirmed. If you're uncertain about a dataset speak to the people or organizations who produced it.

To help build credibility and transparency among your readers it's a good idea to open-source your raw data. This will allow others to build upon and add authority to your work.

Remember:

It's easy to lie with data, but it's even easier to lie without it.

2. THE DATA JOURNALISM PROCESS

Now that you have a basic understanding of what data journalism is, let's have a look at how it is done.

Through our own experience, the best data-driven stories start their lives as a series of questions.

E.g. - "How many people are affected by ...", "Where are the most cases of ….." or "Is this a pattern?"

By focusing on first asking questions and then looking for data to find answers, your story is more likely to have a real impact on people and make them care.

The 4 step data journalism process:

2. GET DATA

Be flexible to adjust parts of your story based on the available data

1. STORYBOARD YOUR IDEA

3. CLEAN AND ANALYSE DATA

Check for missing or incomplete values and look for supporting data at needed

Check for missing or incomplete values and look for supporting data at needed

4. DELIVER INSIGHTS

STEP 1:
Storyboarding

The main ingredient to a successful data story is creativity.

Data by itself is not a story. It requires you to think creatively about what's relevant to your audience and what is not.

On the flip side, a great story idea without data is also not a data-driven story. Often, finding the right balance between what story you want to tell vs. what data you have requires some trial and error.

A mistake a lot of inexperienced data journalists make is thinking that they need to analyse big datasets to tell a story. A better approach is to start off with smaller datasets and develop them over time. This will help develop data-fluency and ensure more effort is placed on extracting the story's meaning.

~~~~~~~~~~~~~~~~~~~~~~~~~~~

What makes a good data-driven story?

1. Contextual and explainer stories

As we've mentioned several times already, data-driven stories should focus on telling a story. To attract larger audiences, these stories should be newsworthy and topical. The most desirable stories weave data within the fabric of the story to enhance the understanding and provide context to a topic or issue.

2. Dense or complex stories

Stories shrouded in numerical complexity are great candidates for data journalists to tackle. With the right analysis and interpretation, data journalists can transform a complicated story into something that is manageable and easy to digest for readers.

3. Exploratory or interactive stories

Interactive stories can help personalise or reignite interest in an ongoing topic. These types of stories involve readers by allowing them to search or filter through a dataset. While giving readers access to raw data can sometimes be helpful, you should always strive to deliver this data within a story narrative.

4. Investigative stories

Many investigative journalism projects involve analysing a large collection of documents or datasets to understand why and how something has happened. Data journalists can play a key role in extracting and analysing this information in order to inform the public.

Remember:

Not every story has to find a hidden meaning in the data, sometimes you can just use data to enrich the understanding or experience of your story.
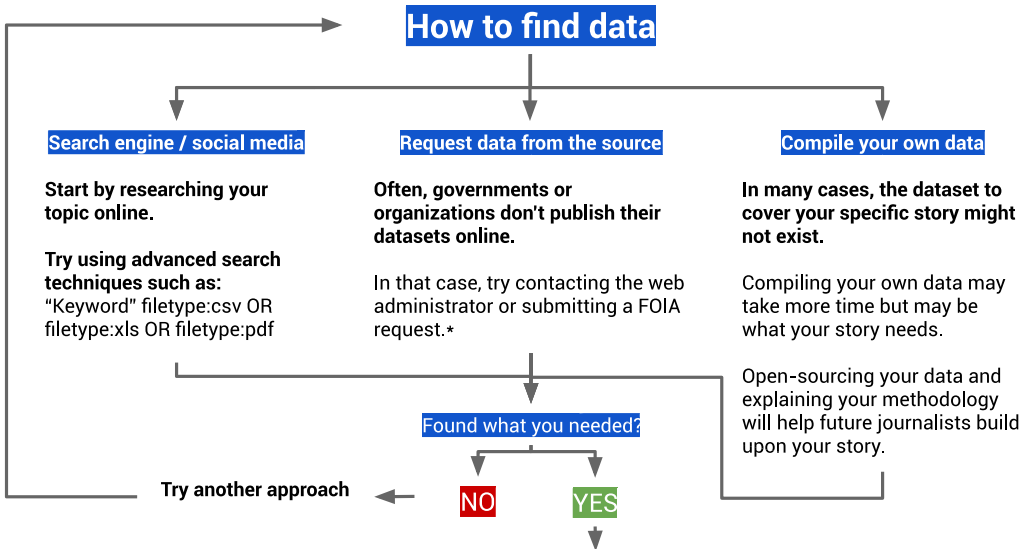
STEP 2:
Get data

Sourcing good data is often cited as the biggest challenge data journalists in the Arab world face today.

While this problem is not unique to Arab journalists, it does create an additional barrier to adopting data-driven reporting within newsrooms. The flowchart gives you some tips on how to find the right data for your story.

How can I find the right data for my story?

How to find data

Search engine / social media

Start by researching your topic online.

Try using advanced search techniques such as:
"Keyword" filetype:csv OR filetype:xls OR filetype:pdf

Request data from the source

Often, governments or organizations don't publish their datasets online.

In that case, try contacting the web administrator or submitting a FOIA request.*

Compile your own data

In many cases, the dataset to cover your specific story might not exist.

Compiling your own data may take more time but may be what your story needs.

Open-sourcing your data and explaining your methodology will help future journalists build upon your story.

Found what you needed?

Try another approach ← NO   YES

Finding the right data and getting it in a format that you can use can be two very different things.

To download data, many websites may offer a downloadable csv, xls or pdf file. Other websites may offer an API for your developer to tap into. Before moving on, make sure you've answered all the questions above about trusting your data and the ethics of using it.

Move onto cleaning and analysing

Freedom of information legislation exists in Jordan, Tunisia and Yemen.

STEP 3:
Clean & analyse

Once you have your data you can begin the process of cleaning and analysing it.

Cleaning data starts with converting it into a format that you can make sense of, for example, extracting tables from a pdf document into a spreadsheet[1].

The next step is to check for incorrect, missing or duplicate values.

Spending additional time thoroughly cleaning a dataset can significantly reduce the chance of drawing the wrong conclusions during your analysis.

How to analyse or "interview" your data?

1. Get to know your data - very well

Analysis involves deriving meaning from your data. To do this, you should ask your data a series of questions such as: What real-world observation is my data measuring? Does it accurately represent my target population? What are the biggest, smallest and average values?

2. Ask critical questions

The quality of a data journalist can be measured by his or her ability to ask a dataset critical questions. Good analysis stems from knowing the subject matter very well and cross-referencing what you expect to happen with what your data is telling you.

---

[1] Tabula.technology, OCR tools

STEP 4:
Deliver your story

The final step in the process is to deliver your story. Remember not all data-driven projects need to be visual.

Choosing your delivery mechanism will depend on what type of data you'd like to present and what skill-sets you have available in your team.
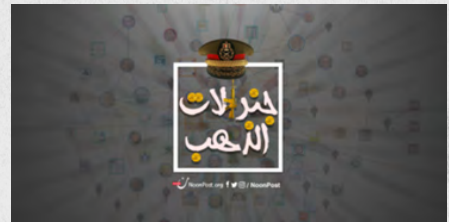
MAP

VIDEO

CHART

ARTICLE

4. EXAMPLES FROM THE ARAB WORLD

In preparation for this chapter, we reached out to several thought leaders from around the region to share with us their experiences in working with data journalism.

Around the Arab World, data journalists from Al Jazeera, Inkyfada, 7iber, Info-Times, Noonpost and the Arij network have produced various forms of data-driven stories.

In the next chapter, we'll be sharing a few case studies from Al Jazeera, Noon Post and 7iber.

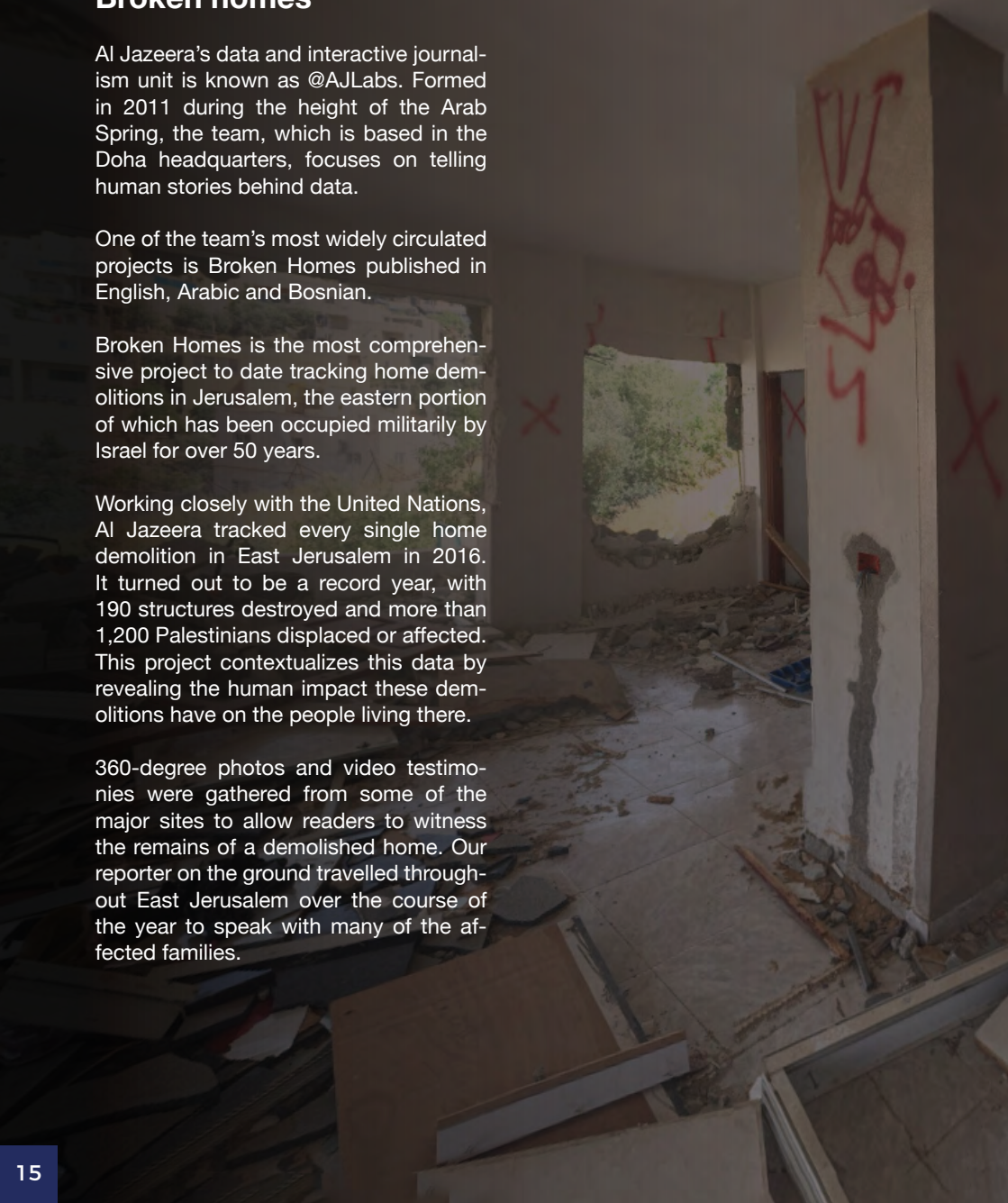# a. Al Jazeera

Broken homes

Al Jazeera's data and interactive journalism unit is known as @AJLabs. Formed in 2011 during the height of the Arab Spring, the team, which is based in the Doha headquarters, focuses on telling human stories behind data.
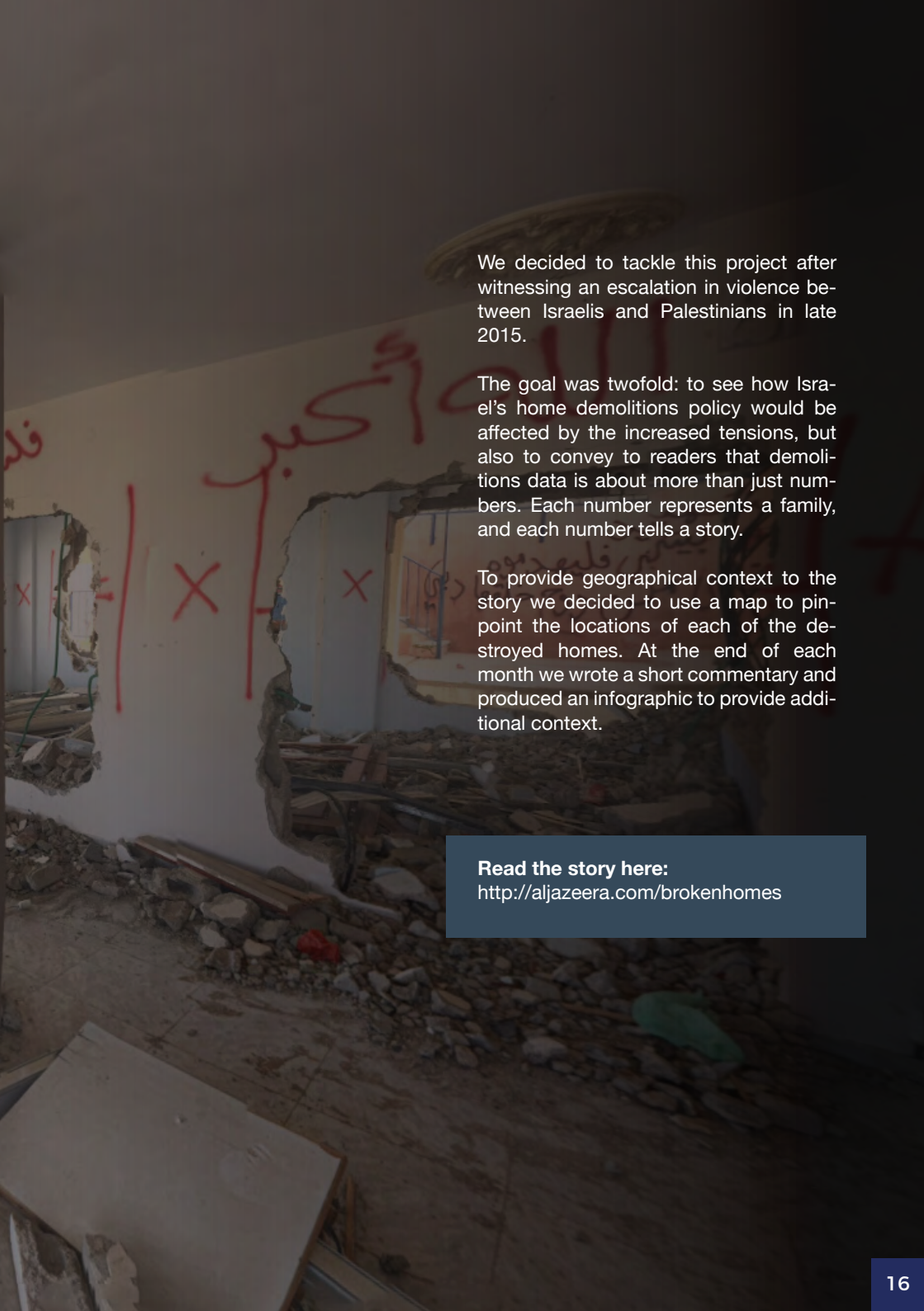
One of the team's most widely circulated projects is Broken Homes published in English, Arabic and Bosnian.

Broken Homes is the most comprehensive project to date tracking home demolitions in Jerusalem, the eastern portion of which has been occupied militarily by Israel for over 50 years.

Working closely with the United Nations, Al Jazeera tracked every single home demolition in East Jerusalem in 2016. It turned out to be a record year, with 190 structures destroyed and more than 1,200 Palestinians displaced or affected. This project contextualizes this data by revealing the human impact these demolitions have on the people living there.

360-degree photos and video testimonies were gathered from some of the major sites to allow readers to witness the remains of a demolished home. Our reporter on the ground travelled throughout East Jerusalem over the course of the year to speak with many of the affected families.

We decided to tackle this project after witnessing an escalation in violence between Israelis and Palestinians in late 2015.

The goal was twofold: to see how Israel's home demolitions policy would be affected by the increased tensions, but also to convey to readers that demolitions data is about more than just numbers. Each number represents a family, and each number tells a story.

To provide geographical context to the story we decided to use a map to pinpoint the locations of each of the destroyed homes. At the end of each month we wrote a short commentary and produced an infographic to provide additional context.

Read the story here:
http://aljazeera.com/brokenhomes

b. Noon Post

The 'Generals of Gold':

In 2017, Noon Post published the Generals of Gold, an ambitious project which aimed at untangling the complex relationship between the Egyptian army and its control of the country's economy. Released over 12 episodes, this project is one the largest Arabic Data Journalism project in the region.

The team used network diagrams and analysis to show the systematic exploitation of the Egyptian economy favoring the ruling elites.
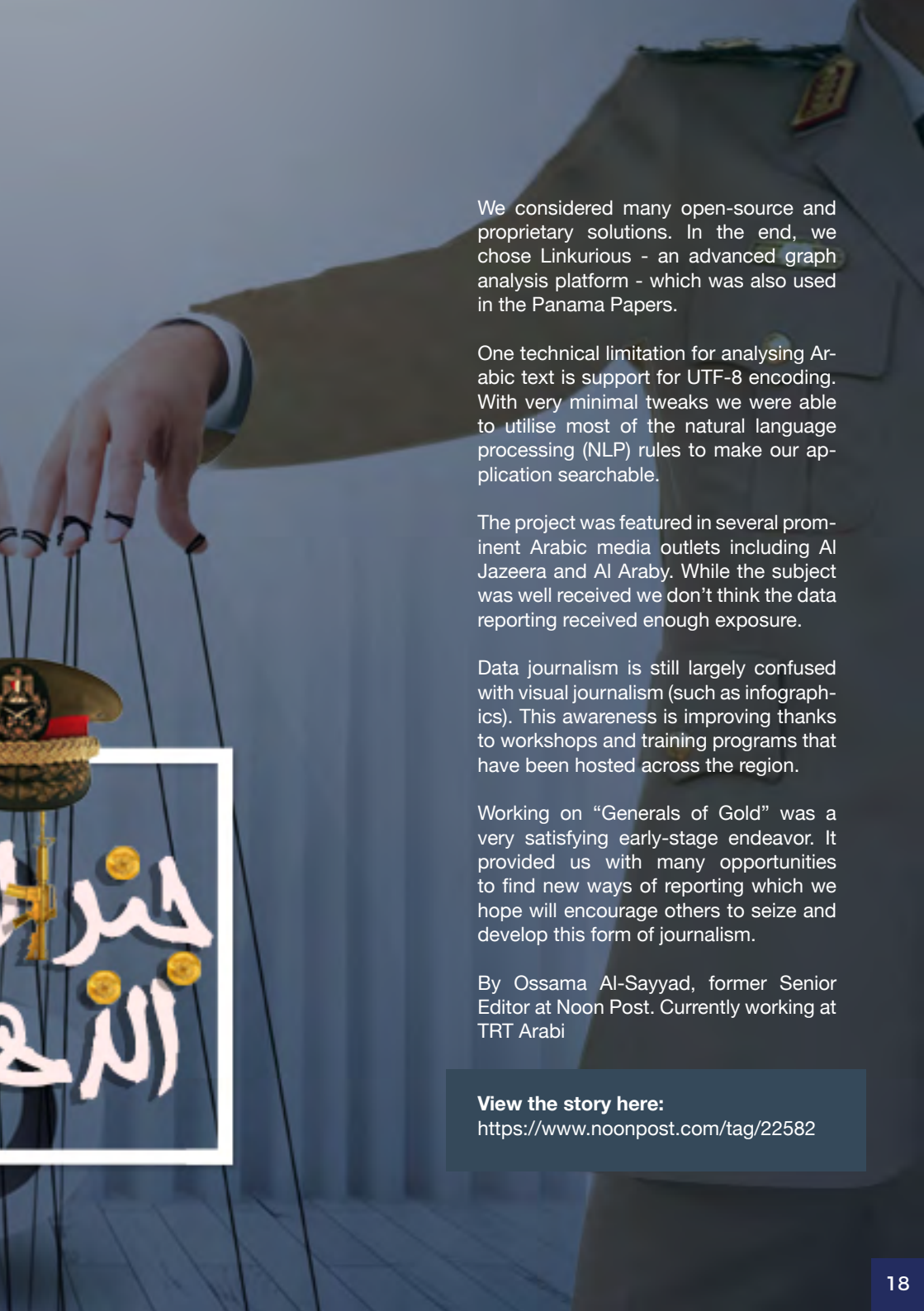
The project allows for any researcher or journalist to access the data and contribute to it. By taking this approach Noon Post was able to transform this story into an ongoing resource.

The project succeeded in making sense of a large amount of data which would have otherwise been neglected in the political and economic scene.

The first big challenge the team faced was sorting through the piles of unstructured data. If this wasn't enough, we knew that each piece of information had to be fact-checked and validated before it could be used.

Prior to this project, no one in the team had any experience with graph analysis or network modeling. We spent a lot of time teaching ourselves this form of analysis and looking for the right tools to use.

We considered many open-source and proprietary solutions. In the end, we chose Linkurious - an advanced graph analysis platform - which was also used in the Panama Papers.

One technical limitation for analysing Arabic text is support for UTF-8 encoding. With very minimal tweaks we were able to utilise most of the natural language processing (NLP) rules to make our application searchable.

The project was featured in several prominent Arabic media outlets including Al Jazeera and Al Araby. While the subject was well received we don't think the data reporting received enough exposure.

Data journalism is still largely confused with visual journalism (such as infographics). This awareness is improving thanks to workshops and training programs that have been hosted across the region.

Working on "Generals of Gold" was a very satisfying early-stage endeavor. It provided us with many opportunities to find new ways of reporting which we hope will encourage others to seize and develop this form of journalism.

By Ossama Al-Sayyad, former Senior Editor at Noon Post. Currently working at TRT Arabi

View the story here:
https://www.noonpost.com/tag/22582

c. 7iber

The Arab world has some good data journalism websites such as Inkyfada, 7iber and InfoTimes, in addition to few investigative journalistic reports produced by the Arij organization.

Yet many factors have interfered with the adoption of data journalism by the Arabic media and these include the lack of data or categorized data, and the inability to make it available because the process often takes a lot of time. Moreover, quick news have become the most common type of content for media companies in the Arab World and there aren't manycompanies that use data as a primary source of content. Many media companies in Jordan for instance monetize from advertising, which imposes a certain type of content, that is quick and time-sensitive, to optimize readership. This forces editors to limit themselves to timely news, without having the luxury of waiting for data to be collected and analysed in order to develop hypotheses, hence narrowing down their field of work.

When discussing the factors that are delaying data journalism development in the Arab World, we must always look at the political side of the problem. The lack of public data is one of the methods authorities use to prevent people from accessing information and building their own hypotheses and arguments. If journalists succeed in overcoming this obstacle, they will learn how to use information and develop new skills, to the extent that some organizations will be forced to regard these skills as a journalistic art that requires prioritization and financial support. Other factors include

www.7iber.com

weak archiving, classification and data numbering, especially from official institutions.

In some Arab countries, the 'right to information' act is often used against journalists. Even though it forces institutions to share information, it also allows them to take more time to respond. This means journalists might end up waiting up to 30 days to get answers and their requests could get rejected. If they do get rejected, the appeal procedures won't be binding unless they turn to court and this can be costly. That said, the right to information act is worthless without the development of confidential laws and the security of state documents.

This is needed because data and information are being classified as confidential without being really linked to state security, so when someone requests information, they get told that the information they're seeking is confidential.

The other problem in data journalism is the inability to differentiate between data visualization and producing journalistic reports based on data analysis.

I believe that some of the stories that succeeded in communicating their hypotheses by using data, have made data journalism an important subject for organizations concerned with investigations and in-depth stories. If these organizations were able to offer financial support, then they should allocate few departments to data journalism work.

By **Dana Jibril**, journalist at 7iber (Jordan)

5. HOW TO GET STARTED

Starting your first data project will require assembling the right combination of journalists and technologists. This often involves breaking down organisational silos and working across departments.

The first step will involve obtaining management's buy-in to the project. From our own experience, the best way to achieve this is to speak with other teams about the kind of work that they do. Your goal should be to bridge the gap between editorial and technology.

a. Bridging the gap between editorial and technology

Your goal should be to bridge the gap between editorial and technology.

Here are the risks involved in focussing too much on a particular approach:
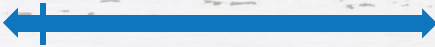
Technology/Design dominated:

Taking a technology or design dominated approach may result in a fancy product but often at the expense of the story. It is recommended to experiment with various forms of designs and technology but always make sure that you're delivering your story in the best way possible for your readers.

Editorial dominated:

On the flip side, too much dependence on traditional editorial processes often results in unimaginative or very text-heavy stories. It is common for news organisations to start their data journalism journeys here. Continuously adopting new storytelling techniques will help develop your newsrooms culture of experimentation and creativity.

Ideal scenario:

Ideally, your editorial, technology and design efforts should all be in balance. Everyone should be working towards delivering the best possible story using the techniques right for your audience.

b. Developing a successful team

We asked two veteran data journalists:

What makes a successful data journalism team?

Currently, data-driven journalism is still looked at as a luxury by the majority of newsrooms and media outlets in the Arab world. Most managers are more concerned with their online presence and content curation, ignoring the fact that riding the digital transformation will optimize these processes.

I believe that this can be changed by working first on the data mindset and nurturing a tolerating environment to transform the working.

It requires communicating with team leaders and hosting data evangelists to show different employees the power of becoming data-driven at all aspects. Also, we need more collaborative initiatives to break the ice between media practitioners and technologists.

At the same time, journalists should tap into the open-source community to help develop the right Arabic language tools and platforms.

- **Bahia Halawi** - Co-founder & Data Engineer at Data Aurora.

Diversity is the key ingredient to make a successful data journalism team. Everyone needs to realise that combining different skill sets creates unique and interesting work. Data journalism can't function in silos, it needs conversation, arguments, complete and utter failure and experimentation.

- **Alia Chughtai** (Online Producer)

Typical team roles:

Most successful data-journalism stories will involve a combination of different roles. This does not necessarily mean four different people. Often times you may have a journalist with design or development skills.

Job role	Responsibility
Data Journalist	The data journalist is responsible for crafting the story. This person is typically trained as a journalist or is a subject specialist who is passionate about telling a story.
Editorial Designer	The designer's role is to produce the functional and aesthetic design to best deliver the story to the audience.
Developer/ Data Analyst/ Data Scientist	The developer/ data analyst or data scientist is responsible for transforming and analysing the data so that it can be understood by the rest of the team.
Data Editor	The data editor makes sense of the complete story and finds the best way possible to deliver the story to the audience.

6. Tools
   and datasets

Tool	Link
Data Extraction	
Tabula - extract data from PDFs	https://tabula.technology/
Document Cloud	https://www.documentcloud.org/
Data Cleaning and Analysis	
Google Spreadsheets	docs.google.com/spreadsheets
Open Refine	http://openrefine.org/
Data Visualisation	
Datawrapper	https://www.datawrapper.de/
Infogram	https://infogram.com/
Flourish	https://flourish.studio/
Data journalism frameworks	
Workbench	http://workbenchdata.com/ (No coding required)
R - Tidyverse suite	https://www.tidyverse.org/ (for programmers)
Python - Pandas	https://pandas.pydata.org/ (for programmers)

Useful data sources:

Name	Type of data	Link
UNDP	Arab Human Development data	http://www.arab-hdr.org
World Bank Open Data	Worldwide demographic and economic data	https://data.worldbank.org/
Local bureau of statistics websites.	Country specific data including health, education, census, population and data on public resources.	E.g. Palestine: http://www.pcbs.gov.ps/default.aspx Qatar: https://www.mdps.gov.qa

About the author:

Mohammed Haddad is a data journalist and instructor. He leads Al Jazeera's interactive team, @AJLabs, where he uses data to tell compelling visual and analytical stories.

He's interested in developing data journalism in the Arab world through regular training sessions and international conference appearances. He is the co-founder of PalestineRemix.com, a multimedia rich website available in four languages (English, Arabic, Turkish and Bosnian) for narrating the Palestinian-Israeli story.

His passion is finding interesting ways to tell human stories hidden within data.

ALJAZEERA
MEDIA INSTITUTE

AJMInstitute

+974 44897666
institute@aljazeera.net
http://institute.aljazeera.net/en