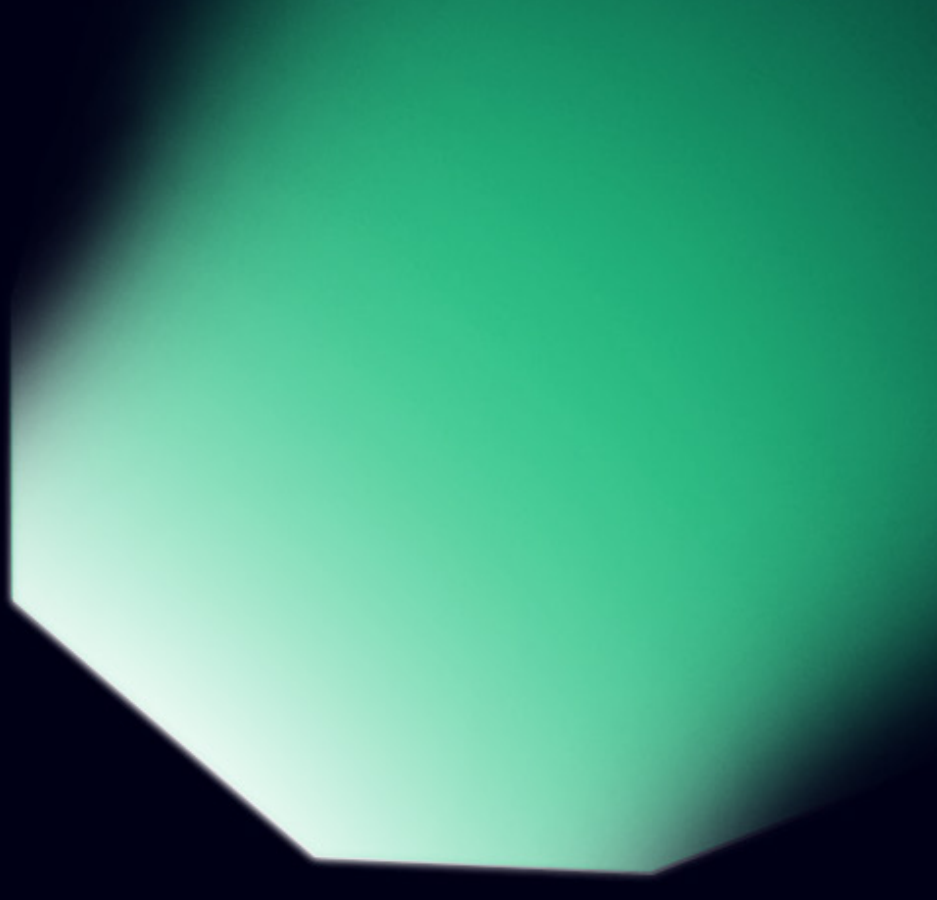


OFFSHORE JOURNALISM

A project to maximize free speech by exploiting different jurisdictions



FINAL REPORT - MAY 2018

About

The Offshore Journalism Toolkit was a project that ran in 2016 and 2017 with the aim of helping publishers store their content beyond the reach of censorship. The project created a process to this end, as well as the tools to implement it (available on Github at <https://github.com/jplusplus/CrawlerToolkit/>) and research which are presented in this report.

Credits



The Offshore Journalism Toolkit is a project by Mario Tedeschini-Lalli and Nicolas Kayser-Bril, with help by Anne-Lise Bouyer, Pierre Romera and Defne Altioek. It was financed by *Google's Digital News Initiative*.

Official website: www.offshorejournalism.com

The final report has been curated by Donata Columbro, with help by Serena Carta. Graphic design by Alessandro Rizzo

OFFSHORE JOURNALISM

Final report

May 2018

Index

<i>Contribution</i>	The story of the Offshore Journalism Toolkit <i>Nicolas Kayser-Bril</i>	05
<i>Contribution</i>	Publishing for the future and freedom of the press <i>Mario Tedeschini-Lalli</i>	07
<i>Interview</i>	Digital preservation of online news: challenges and best practices <i>William Kilbride</i>	10
<i>Interview</i>	Freedom of expression in the digital era: how to preserve and enhance it? <i>Guy Berger</i>	13
<i>Interview</i>	Digital preservation in the art sector: a different point of view <i>Dragan Espenshield</i>	16
<i>Contribution</i>	Metadata advantages: from digital preservation to fact checking <i>Claudio Agosti</i>	19
<i>Focus</i>	Right to be forgotten and freedom of expression: in search of balance <i>Ernesto Belisario</i>	22
<i>Resources</i>	Tools for web pages and documents filing <i>Andrea Borruso</i>	24
	Authors and contributors	29

The story of the Offshore Journalism Toolkit

5

Nicolas Kayser-Bril

Data-driven journalist.

One of the authors of the Offshore Journalism Toolkit.

One day of July 2016, Mario Tedeschini-Lalli, head of innovation at Gruppo L'Espresso, an Italian media giant, published a blog post where he pointed out that much of the work we - online journalists - do will probably be lost forever if it is not archived. I joked with him on Twitter that we should apply for funding to Google's Digital News Initiative, were it only to put down in writing what should be done to preserve online content. Our application was tongue-in-cheek from beginning to the end.

Our slogan was "corporations use offshore vehicles to optimize pro-

fits, publishers should do the same to optimize freedom of expression". We even had a budget line for a €3,000 study trip in Iceland (we didn't do it).

“**Our slogan was “corporations use offshore vehicles to optimize profits, publishers should do the same to optimize freedom of expression.**”

We were stunned, a few months later, when Google actually gave us €50,000 to build the Offshore Journalism Toolkit. We first mapped the situation.

We called publishers to know when and in what circumstances their content was permanently removed from their servers.

We called archivists to know what public libraries were doing to pre-

CONTRIBUTION

serve content for the next generations. We called lawyers to understand what publishers could do if a court ruled that an article needed to be removed. And we examined in-depth several instances where, in Europe, publishers had had to delete online content.

The results of several dozens interviews were very promising. It turned out that content deletion was much more widespread than we thought. Dozens or hundreds of news articles are being deleted, each month, from the archives of online news outlets. Some organizations were even bought by rich men for the sake of destroying their archive (most famously Gawker).

All the while, national libraries were archiving content, but were doing very little.

The idea of preserving interactive or video content, for instance, was still a plan for the far away future. And some cases were simply too absurd to be believed. The French censorship authority, for instance, is checked after the fact by a single civil servant from another body.

But this lone civil servant is simply given a list of URLs that were censored - and cannot access the web pages in question because they are blocked!

We set out to work. Our plan was to create a standard for an HTML meta element (a line of code written on top of a web page) that publishers could insert in their articles. We then built a simple robot, that could be set up by anyone, to crawl the web looking for the meta element in question and, each time it found a page containing it, archive it on different servers in several

countries. Once stored safely, the article wouldn't have been at risk if the original were to be deleted.

The beauty of this design lied in its legal leanness, for publishers didn't have to break the law to save their content (if a publisher archived an article after a judge ordered it censored, it would be contempt of court).

It was also extremely cost effective. Publishers just had to add a line to their web pages and anyone could then, for a few cents a month, run a crawler. It was a decentralized search-and-rescue network for journalistic content.

“***Publishers do not have an incentive to preserve content for future generations. After all, their job is to publish, not archive.***”

We tested the set-up with L'Espresso and PrimaDaNoi in Italy. Although it worked well, technically speaking, it did not take off. Publishers were very sympathetic to the idea of preserving content but, when we asked them to try out our system, very few went ahead. Probably because they saw it as an additional cost which, as low as it may be, would have made their already precarious situation worse.

Publishers do not have an incentive to preserve content for future generations. After all, their job is to publish, not archive. Our experiment is still useful. We now know that we must leverage other stakeholders, such as journalists and readers, to save our cultural heritage.

Publishing for the future and freedom of the press

7

Mario Tedeschini-Lalli

| Journalist and a consultant on Digital Editorial strategies. One of the authors of the Offshore Journalism Toolkit.

The Offshore Journalism Toolkit emerged as a possible way to solve the problem of the increasing amount of accurate and legally published journalism that is being deleted by order of local authorities, thus preventing future generations to access it. The idea was to attach a meta tag to any piece at risk, making it available for copying by organizations in more liberal jurisdictions. As the project comes to an end we can point to at least a couple of more general problems that arose.

- One is a cultural, maybe also financial problem in the industry. As Nicolas Kayser-Bril put it, “publishers do not have an incentive to preserve content for future generations. After all, their job is to publish, not archive”.
 - Another one has to do with the basic assumption of our project that it would be possible to exploit different jurisdictions to maximise freedom of speech, the existence of which at present and in the future seems now less certain.
- Our approach was mainly publisher-centric, but the cultural problem can and should be addressed first of all by journalists, if they are not on board it would be difficult that publishers would.
- Although there are already a number of tools that let journalists save their own content,¹ the main issue here is to understand that “publishing” and “archiving” in the digital world is actually the same thing, or - as we pointed out in our report - that the potential “currentness” of any item

¹Lauren Hazard Owen, *Here are three tools that help digital journalists save their work in case a site shuts down*, IJNET, November 23, 2017

CONTRIBUTION

digitally published in the past is a feature that expands across time our ability to inform and influence societies. We should all embrace the idea that “freedom of the press in the digital environment includes the freedom of ‘publishing for the future’, and that limiting or curtailing it puts freedom of speech at risk”.²

We see, instead, a creeping acceptance of the more historically limited concept of freedom of the press and of freedom of expression, rooted and stuck in our pre-digital past.

Of course, “publishing for the future” entails new responsibilities, as well as new opportunities. New correction and updating policies, for instance, should be discussed and implemented with the same degree of attention, human and technical resources assigned to “current publishing”. Overall, we think that an industry-wide dialogue about this issue involving all stakeholders should be organized, especially at the European level.

It could be an on-going, transparent process to try and re-define values, offer new criteria and devise new tools. The first step could be a call to discuss guidelines on erasure, editing or updating published content, which a news organizations could then make public.

As far as the jurisdictional assumption is concerned, many signs point to a growing nationalization of the Internet. On the one hand authoritarian regimes appear to be growing in number and in their power to muzzle free speech (even

within the European Union), on the other hand even in more liberal societies speech control seems to gain momentum, be it to allegedly combat terrorism, hate, or violence in general, or to guarantee the citizens’ privacy in the digital environment. The even louder call to regulate digital platforms and their power in these respects, may have the consequence of making the platforms the extra-judicial arbiters of free speech, which - coupled with the requested accountability to national governments - may push them to limit their use as a means to circumvent censorship by local authorities. The most recent instance being the announcement by Signal, an encrypted messaging app, that Google and Amazon stopped letting the app use “domain fronting” within their cloud service, which is a way to make it available in countries that block it.³

Although we are not sure how, or if this trend can be successfully reversed, we think that the news industry as a whole should at least be aware of these risks,

and not fall in the trap of thinking that any platform-bashing is in the end a good thing for journalists and news publishers. Of course platforms have their responsibilities, and they could use much more oversight by citizens as well as rules to “help” them in terms of transparency and competition, but free speech should be the ultimate civic test on such measures - journalism, let’s not forget it, has a vested interest in free speech.

If and when journalists and publishers become aware of the kind of problems we tried to describe, they may come together with other sta-

“**The potential “currentness” of any item digitally published in the past is a feature that expands across time our ability to inform and influence societies.**”

² Nicolas Kayser Bril and Mario Tedeschini Lalli, *Offshore Journalism. Preliminary report*, June 2017

³ Abrar Al-Heete, *Signal says Amazon, Google will no longer help it evade censorship*, CNet.com, May 1, 2018

CONTRIBUTION

keholders (human rights organizations, digital platforms, tech groups, etc.) to set up a distributed “Search & Rescue” service to save content at risk, thus making more difficult for governments to control journalistic content. The “stress signal” and its mechanism may be similar to the one we proposed or a different one, but we do believe such a service should exist.

“*Free speech should be the ultimate civic test on such measures - journalism, let’s not forget it, has a vested interest in free speech.*”

Digital preservation of online news: challenges and best practices

Interview to William Kilbride

| Executive Director of the Digital Preservation Coalition

What is Digital Preservation Coalition's mission? What kind of matters do you address to and why?

The DPC is first of all a membership organization. It exists because a group of organizations got together to share the challenge of looking after the digital data both in a long and a short term.

The coalition started in 2002 as a collaboration between a relatively small number of organizations in UK and Ireland, coming from the public sector and the so called “memory institutes”: uni-

versities, libraries, national archives, museums. All types of organizations that have a challenge in locking material for the long term because preservation is what they do.

Over the years, the number and the types of organizations getting involved in the problem of digital preservation have grown and diversified, and the DPC has reached out the international level. Today the Coalition helps everyone with the need to preserve and save data for long term, from commercial banks to nuclear industry, from universities and libraries to business companies.

INTERVIEW

Our mission is to help them to be able to secure for themselves digital assets for the long term, so that they can survive through different changes and technologies; but also to let digital resources become a robust and useful part of their corporate existence, their cultural memory, whatever the function they intend to serve can be supported with digital material.

We achieve our aim through six different kind of activities: we do work around advocacy, we do quite a lot of community engagement work, we help with training or workforce development, in capacity building, in developing model standards and good practices, we try to build sustainable platforms for good governance.

What kind of challenges do you see in the digital preservation of online news?

I see three main challenges.

Obsolescence. Typically, in the early history of the Coalition and digital preservation, there was a lot of focus on matters related with obsolescence - meaning what things you are going to do to make sure that your digital objects are protected. That became obvious over the years and it is still a big challenge.

Political interference. But what became obvious over the years as well is that there are other types of risks. So, last year we did a thing called “the bet list”: we wanted to make a list of digital contents in danger to be seized - like the global red list of species in danger of extinction! The

things on the list were not what you would have expected to be if all you were concerned about was obsolescence. The nominations of the list included things like environmental data from US governmental scientific research. Now, it seems to me that US environmental data is suddenly at risk because Donald Trump is the President of the United States and he has perhaps a different view on climate science than the scientific community. So, the point here is that there are significant risks to the digitally world, because material can be easily deleted, easily falsified

and easily lost. Another good example of this trend comes from a study that shows that in 2009, two years after Tony Blair’s having left office in UK, 40% of the links to web resources giving answers to Parliamentary requests and website citing evidence for public policy in the UK were broken and no longer available. That is a significant problem, because these were no random websites but answers to Parliament requests and links to important matters of public discourse! So, I think that we need an authentic legitimate record of what our politicians have said and done and this is something that digital

preservation can help us to do, to not overlook the fundamental servers.

Big companies. There is also a business issue here, a financial one we have seen growing in last years: sustainability. We see great services, bubbling up, becoming very popular for a short while and then disappearing. I think that the

“ I think that we need an authentic legitimate record of what our politicians have said and done and this is something that digital preservation can help us to do, to not overlook the fundamental servers. ”

INTERVIEW

reason why they are disappearing isn't because of technological problems, but because the company behind them has made business decisions to close servers, and there is not much we can do about that.

I think that what we need from some of the big tech companies like Facebook or Google is either some commitment to preserve and to transparent decision making about how and when the servers are getting off; but they also need to pay their taxes because national libraries and national archives need public resources to go on in their historical mission to record and preserve the public cultural memory.

From your experience, what good practices in digital preservation at international level would you point out?

I can struggle with this question because there are too few good examples! The classic example of someone who is doing a great job on web archiving has been for long Internet Archive. They are member of our Coalition and I won't say anything bad about them, but you have to remember that they have a weakness: they are for nonprofit and they don't have a legal mandate to do this.

They are effectively breaking the law every time they take a copy of a website, because they have no copyright permission to keep this stuff. The big legal challenge against them is the memory vs corporation's interests. So, if someone says something really truthful but that damages a big

corporation and it ends on Internet Archive, it is possible that Internet Archive is going to be sued and they won't simply be able to sustain that.

Speaking about the legal mandate takes me into the world of legal deposit libraries: there is quite a good network of them internationally that are taking a variety of copies of entire domains.

This is also a good piece of work and they are doing it under a legal framework. Their weakness here is that it's very difficult to access that material. In Scotland, for example, the National Library has a really great collection but you can see it if you only physically go there. And, of course, that's a problem.

There is also a very interesting work done by social science research.

I can mention Politwoops, a service able to capture deleted tweets from politicians. More generally, in a variety of academic disciplines there are really active communities in digital preservation; they really know what they are doing but the expertise from those relatively

small groups has not really reached to the broader journalistic or policy making worlds.

**“
What we need from some of the big tech companies like Facebook or Google is either some commitment to preserve and to transparent decision making about how and when the servers are getting off.**

Freedom of expression in the digital era: how to preserve and enhance it?

Interview to Guy Berger

| *Director of UNESCO's Division of Freedom of Expression and Media Development*

How serious is the threat of the right to be forgotten for the freedom of expression and the access to information in Europe?

From the UNESCO's point of view, we interpret the right to freedom of expression as the right to seek and receive and the right to impart information. In the case of the "right to be forgotten" – meaning a right to be delisted as per the European Court of Justice decision, the right to *impart* is not directly limited as the information is

not required to be removed at source. So the impact is primarily on the right to *seek and receive*, because when a content is delisted then it becomes unfindable.

In other words, a right to be delisted is not so much an act of censorship of the right to express, but rather an intervention that limits the distribution of information and its discoverability.

Of course, the two dimensions of the full right have an interdependence; restraints on one do affect the other.

Certainly, it undermines the exercise of the right to meaningful express information, if no one can find that expression. But the immediate impact is on the right to seek and receive.

The issue then is to what extent a restriction such as delisting can be legitimate. In the European law, certain criteria have been set out as justifiable within the decision of the Court of Justice. In terms of international standards, there can be right to be delisted or even to have content removed entirely, depending on the applicability of international standards of legality, proportionality, necessity and legitimate purpose for any restrictions to be considered justifiable. In this wider perspective, restrictions can indeed be justified – such as in terms of the right to reputation versus the right to seek and receive information.

For example, if you are victim of so-called revenge porn, you want and deserve to have it at least delisted if not also taken down.

On the other hand, if you are a politician or business person and you want to be delisted concerning your link to corruption, then the law shouldn't work in your favour. Public interest in free flow of information, including its discoverability, trumps a right to reputation in terms of findability in this latter case.

This matter of delisting needs to be very much independently monitored to assess the extent to which it is occurring in terms of international standards. People do need to have their right of reputation protected – but in a way that is not unjustifiably at the expense of the right of other to seek and receive information.

What is your call to European authorities?

Making the delisting in response to every demand could really lead to a limitation of availability and findability of legitimate information.

It would not be about the right to reputation in many cases, but about violating the rights of individuals and the public to impart, seek and receive information as the basic norm.

Currently, Google does not accept everybody's wish to be delisted - they make investigations and verifications, but are required to follow the European Court of Justice's law when someone successfully contests a refusal by the company to delist certain links.

However, there is also a view by some countries in Europe that they have a responsibility to protect a citizen's rights to reputation even at the global level, i.e. that delisting should be made by Google globally. In other words, that it is not enough for Google to geo-customise its delisting, but should do this universally. This implies extending European jurisdiction to other jurisdictions. If this is pushed to the extreme, if Google is required to delist on all its presences on the basis of everybody's request in every jurisdiction, this

“
People do need to have their right of reputation protected – but in a way that is not unjustifiably at the expense of the right of others to seek and receive information.

could radically reduce the utility of the internet.

In such a situation, the company would likely be pushed to automatically delist as a default stance, in response to any complaint, anywhere, because the burden of investigating would be overwhelming.

In this scenario, the nuance required for balancing rights would be eliminated – a swathe of links would simply disappear from the global internet irrespective of merits.

In light of all that, I would encourage European authorities to keep in mind the dangers of applying one region's jurisdiction on a global basis. This could come back to damage the availability of information of people in Europe and of other countries in other jurisdictions.

In addition, European authorities should look to international standards and the principles of proportionality and necessity in the way that the Court of Justice's law is implemented. More creative solutions should be explored – for instance, instead of delisting globally, a company could be urged to signal for instance that the specific link under contestation is delisted within European jurisdiction.

This “co-listing” approach could be easier to accommodate in terms of greater alignment to international standards of necessity and proportionality.

UNESCO's project to define the Internet Universality concept includes four fundamental principles: R – that the internet is based on human Rights / O – that it is Open /A – that it should be Accessible to all / M – that it is nurtured by Multi stakeholder participation. How they can help the media to preserve and enhance the free movement of ideas?

The ROAM elements are all relevant here. In particular, concerning the so-called European right to be forgotten, this issue should be approached within the recognised universal Rights framework (such as reputation versus expression). There are also risks to Openness and transparency (such as algorithmically driven take-down of links in response to complaints). On the other hand, Openness in the economic sense can enable competition between discovery engines, which can mitigate the risks of a single big actor being the central interpreter of the Court of Justice's decision. Accessibility is impacted, in that there is a need to enhance users' Media & Information Literacy to

understand the rationales and the risks of a right to be delisted. Lastly, because of the importance of upholding the right to expression in any balance with delisting, it is essential that journalists, publishers and editors should be involved in a multi-stakeholder dialogue when it comes to developing specific laws and policy at a country level.

**“
If Google is
required to delist
on all its presences
on the basis of
everybody's request
in every jurisdiction,
this could radically
reduce the utility of
the internet.**

Digital preservation in the art sector: a different point of view

Interview to Dragan Espenshield

| *Rhizome's Preservation Director*

Digital preservation in the art sector: how does it work and what kind of solutions have been identified?

Let's start by saying that digital preservation has been led by the library and archive fields: they have developed a lot of practices that the art world is trying to adapt with the aim to create its own strategies. Secondly, we need to find an agreement on what digital art is: there is the Instagram type of digital art - where a picture can be printed out and bought by a collector - which is not what we are interested in Rhizome. For us digital art is contemporary art engaged with digi-

tal technologies and the internet: we are focused on art that has been created for the internet and makes use of internet and software. The challenge in the arts right now is to have a say to define object boundaries. Many of the works we are dealing with are very specific to the technical cultural environment which are exchanged very quickly.

At Rhizome we work with Webrecorder, a tool we have developed, to preserve websites and we are using several other techniques like server containers or the emulation framework EaaS* to run legacy software in legacy environments.

*Emulation as a strategy for digital preservation is about to become an accepted technology for memory institutions as a method for coping a large variety of complex digital objects (<http://eas.uni-freiburg.de/>).

Can you mention a best practice in the art digital preservation?

We have found that the web archiving approach that we have developed with Webrecorder is extremely productive and produces stable art effects even with technically complex websites that you don't have access to on the back-end side.

For what concerns software preservation, we are working with emulation as a service (EaaS) which is a project we have been involved in thanks to several research programs and it helps us to preserve software really well. Online material and software appear to be boundless and infinite and it seems hard to conceptualize the object that you want to preserve; so, the definition of an object and its boundaries has been the main focus for us. Lot of this is really a curatorial decision: you have to draw those boundaries artificially.

What are the main challenges to be faced in the future?

I think that the challenges I am facing and that software preservation is facing is the whole computing culture, that is moving away from protocols to products. For example, emails have an established and solid protocol that works and is based on open standards that you can go into.

On the contrary, very popular systems like iOS or Android are so lock down that it would be super difficult in the future to reproduce what they have

done. Even if you have a running copy of Android, it might not help you because the application that you might also have a copy of is not much of an application; as a matter of fact, all the computing is happening on a remote server and you can't get hold to that remote server, you never know what the content are because it is proprietary and because of business secrets.

So, what we are trying to develop are ways to void these black blocks systems and understand for example how you can capture network traffic and combine everything locally running software and reproduce behaviors of the software. But you will never be able to capture the completeness of a system that works like that.

Speaking again about challenges, I would like to point out a feedback to The Offshore Journalism project. Strangely in the arts we don't have the issue that we need to remove things: this is something that has been interiorized by the art world, there is a lot of freedom there.

Actually, it is very hard to fight for the deletion of art; if something has entered the museums circle it is pretty safe. Some artists have worked around this idea, creating an exhibition out of medical records, to show what the position of the art world is.

That is also a kind of offshore approach, isn't it?

“ I think that the challenges I am facing and that software preservation is facing is the whole computing culture, that is moving away from protocols to products. ”

INTERVIEW

Tell us something more about your work at Rhizome.

I am directing the Preservation Program which includes research and tool development, but also the preservation of art pieces.

Our own goal is to support artistic practices that engage technology, helping artists who want to engage with it in creating this type of art and institutions that owns these art pieces.

Digital preservation plays a big role in all of it. Actually, not many things have been figured out in digital preservation. How digital preservation usually works is on the term of computer, meaning that you need to conceptualize any system, any object in the running environments where the computer is actually turned on. That is the research focus that we have.

“

In the arts we don't have the issue that we need to remove things: this is something that has been interiorized by the art world, there is a lot of freedom there.

Metadata advantages: from digital preservation to fact checking

Claudio Agosti

Vice president of the Hermes Center, member of the Good Technology Collective

The internet naturally upsets every system used in the past to enter, distribute or file information. We all, contemporary consumers and information producers, especially if we consider ourselves as “innovation witnesses”, have the duty to impose a political agenda that doesn’t follow a profit mindset but that evaluates technology potentials with a view to public interest.

I find it fascinating and puzzling how history has been kept in the past and how people have created stories to acquit themselves, rewrite and

reinterpret facts. Nowadays this action has been digitalized – and I’m wondering if it is becoming stronger and if it will be entirely in Google’s hands (and no more in those of the Roman Catholic Church!). Thanks to this article, let’s now try to picture history, and this history is owned by mankind, and the only useful goal for society is an accessible, long-lasting filing that is resistant to changes of power.

Let’s now split this goal in smaller components. The first one is the “robust indexing” that in the digital world is made with proper metadata, i.e.

CONTRIBUTION

data that describe data. Nowadays a piece of news about a corruption scandal within a public institution lasts 24 hours; metadata that describe the institution, the crime, the amount of money, the context, the victims are all eternal information. Creating correct or wrong metadata is something that we try to do when we have to file more than a hundred photos. The skill to imagine index and tags enduring the passing of time and representing a usable value in the future is a challenge that coexists with the reliability of sources of information.

I believe that this is a metadata problem too: those who produce information should describe the context in a way to allow third parties to revise it easily. I will produce a parallelism with computing, in order to better explain what I mean. In computer security, you need to formalize interactions between software and users: in this way you become aware of the responsibility chain and you can verify the information flow.

For us engineers these are inputs, for a journalist they are information to be verified. The core is that the issue of algorithm gatekeeper and misinformation within social networks could take inspiration from the history of computer conflicts.

One of the most frequent attacks on the web is phishing. The attacker sends an e-mail that embodies a trusted reality, such as a bank, trying to arouse an action by the receiver; if the user is taken in, the attacker may steal him/her something. Therefore, we should consider misinformation

like a form of phishing: you can be victim without taking action but just by reading.

To continue this similarity, let's think about computer networks of the 80s that trusted only known contacts. In information society, this means reading only reliable sources. The suffering publishing industry of the last century would have liked this approach but it would be a disaster for citizen journalism and for independent bloggers.

The user exposed to misinformation is therefore comparable to a computer exposed to cyber-attacks. Computers have raised the number of indicators they used to assess a connection, evaluating the source (that in this similarity is represented by the publisher and the newspaper) and running content and context analysis. In this way we can have connections to unknown devices and stop them if we think they are unsafe. The idea is that if we could have a basket of additional information in each article, we as readers could do this selection too, alone or trust in the collaborative filter of the people we trust.

What does validate a piece of news? The verification of sources, context understanding, cross searches. These are metadata. Why aren't they transmitted and formalized? This wouldn't mean disclosing sources but, as far as possible, allowing readers to repeat the verification procedure. And in case of misinformation or false metadata, then the user could apply a ban as it happens between computers.

“
The skill to imagine index and tags enduring the passing of time and representing a usable value in the future is a challenge that coexists with the reliability of sources of information.

CONTRIBUTION

The formalization of sources is carried out with a metadata structure; our applications would choose according to our instruction, and our priorities would be respected instead of undergoing algorithms with no neutral intelligence that try to imagine what counts for us.

Metadata are organized and planned in a lasting and reusable logic in order to identify different components inside the news. Where there is a historical assumption, there will be a link to the past, where there's an opinion, there will be the author's name and maybe his/her sources. You'll have to tell if someone has validated photos and who is their author. In short, the more metadata there are, the bigger will be the value coming from the information.

The journalist himself/herself has to learn to produce metadata: this would formalize even the verification effort that otherwise on digital media would go undetected. This happens because too often editorial products seem a box of text surrounded by advertisement as every other blog, portal or post that circulate on the web. The aim is to be able to give value and public acknowledgement to this validation, creating new business models based on metadata use in the publishing industry.

Right to be forgotten and freedom of expression: in search of balance

Ernesto Belisario

E-Lex law firm lawyer, member of the Italian Government's Board on Innovation and Digital Agenda

22

Right to be forgotten. This is how often we define the 'right to oblivion', yet it has now become a limiting definition.

The right to be forgotten is indeed one of the most important people's right in today's society of information.

While it can be confused with the right to personal identity (i.e.: the right that safeguards the public image of an individual) this right in fact concerns the safeguard of privacy of a person.

Initially this was meant as right to prevent the re-disclosure of news that are already of public domain after a long time; yet in web and social media era, being forgotten means to obtain the elimination of personal data when they are not relevant or no longer needed for the purposes for which they were collected in the first place.

In the online world, the right to be forgotten can be compromised not only by a re-publication of a piece of information, but also by the fact that such information just remains online.

On the other hand, by making sure information is removed, we allow for the individual's specific news and events – when there is no conflict of public interest, to be forgotten (or rather, to be no longer associated with such person). The right to be forgotten – after it was recognised by law⁴ – has its foundations in the European legislation on personal data protection,⁵ that by art. 17 establishes that the interested party has the right to obtain the cancellation of those details relevant to him or her without any further, unjustified delay.

According to such rule, necessity, proportionality, relevance and surplus of information must be complied with and guaranteed not only by newspapers but also by anyone publishing personal data online (on blogs, forums, etc.).

From a practical standpoint, the application of this new right becomes an answer to the non-trivial question of after how much time it is possible to claim the right to be forgotten? The lawmakers do not define a unique and absolute time limit,

⁴It should be noted, in particular the sentence of the Court of Justice, dated 13th May 2014, lawsuit C-131/12 (also referred to as "Google Spain").

⁵EU Regulation 2016/679 of the Parliament and of the Council on the safeguard of personal data (referred to as GDPR).

so each instance needs to be assessed on a case-by-case basis. Now, as there is often disagreement between those who claim the right and those who manage newspapers and news sites, this can lead to legal disputes.

Knowingly, one of the prerequisites for a piece of news to continue to be published is that there is a public interest in it. For example, if corruption accusations involving authorities arise, the public opinion has an interest in knowing who are the parties involved, what are the operation details and all the legal implications.

This means that the public interest in the events will last over time. Yet, after some time has passed since the events and no updates are given, such public interest ceases, hence prerequisites to claim the right to be forgotten arise.

That's exactly when it becomes difficult to establish a real balance between the right to be forgotten and the freedom of expression and speech. In Italy the Supreme Court of Appeal⁶ has accepted the right to be forgotten for a politician who – under allegations of corruption – had been acquitted at trial and was complaining that after so many years the news was still public on the site of the newspaper in question.

The same Supreme Court of Appeal,⁷ – this time subject to serious criticism -- declared that the right to be forgotten may be exercised when the trial is still ongoing and only after two years since the facts occurred, somewhat restricting – potentially, a lot – the right to report news.

⁶ Supreme Court of Appeal, III Civil Division, 5th April 2012, no. 5525.

⁷ Court of Cassation, Sec. I Civ., 24th June 2016, no. 13161.

Tools for web pages and documents filing

Andrea Borruso

| *President of onData, a nonprofit organization focused on open data, civic technology and investigative journalism*

Find below a short list of tools and resources to file web pages and documents. Some of these are user-friendly, easy for everyone to use immediately; others may be suitable for expert users with some theoretical and/or instrumental knowledge.

Internet Archive (IA)

Internet Archive (<https://archive.org/>) is a not-for-profit digital library aiming to allow an ‘universal access to knowledge’⁸. Along with Google cache, this is the most used online space to search for a copy of a no longer existing or previous versions of web pages. But that’s not all: this tool allows you to actively contribute to building the digital archive itself. Here are some ways to do it.

Basic mode

This is the simplest mode: just open the ‘Wayback Machine’ homepage section (<https://archive.org/web/>), enter the URL of the page you want to file and click on ‘Save page’.

If the administrator of the website has not blocked the access to the crawlers,⁹ the selected page will be filed.

 **Save Page Now**

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

⁸https://www.wikiwand.com/it/Internet_Archive

⁹<https://www.wikiwand.com/it/Crawler>

Browser extensions

Wayback Machine

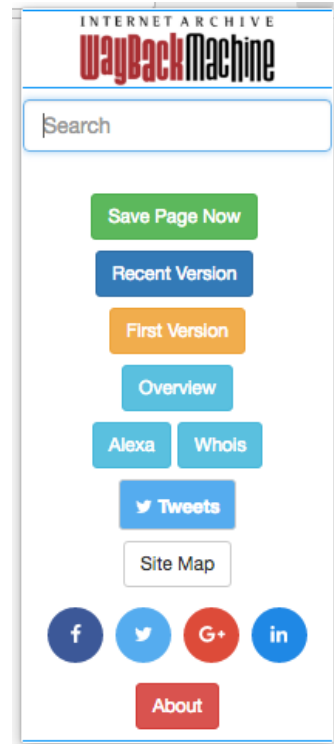
This is a really useful mode: just install an extension (available both for [Firefox](#) and [Chrome](#)) to get a button on your browser that allows you to save the current page, view its most recent version or even the first version already filed in the archive.

Bookmarklet

A bookmarklet is a JavaScript program that can be saved in the browser as a bookmark.

Wikipedia has created this feature for Internet Archive: you just have to drag & drop into your bookmarks and then click it every time you want to save the open page on your browser.

In the same page, you can find a bookmarklet, to view what you have already filed and another to use when a page no longer exist (dead page).



Upload form

This is an online form to upload one or more files (documents, videos, audio, etc.) you are entitled to share on Internet Archive.

You need to create a user account from this page <https://archive.org/create/> and then upload the files you want to archive.

You can enter into the form a set of metadata (added information on the files you are uploading), such as a description, keywords (tag), content creator, date of creation, language etc.

These details will make the search for the uploaded files on your Internet Archive easier and more efficient.

Page Title * 1481564845_7fa3369d02cc06c6d73542e4e 0e 9f 26c ✓	Drag and Drop More Files Here or Select files to add				
Page URL * https://archive.org/details/1481564845_7fa3369d02cc06c6d73542e4e... ✓	<table border="1"><thead><tr><th>Items</th><th>Size</th></tr></thead><tbody><tr><td>1481564845_7fa3369d02cc06c6d73542e4e9G6c.p7m</td><td>29 KB</td></tr></tbody></table>	Items	Size	1481564845_7fa3369d02cc06c6d73542e4e9G6c.p7m	29 KB
Items	Size				
1481564845_7fa3369d02cc06c6d73542e4e9G6c.p7m	29 KB				
Description * Add a description of the item page ✓					
Subject Tags * Add keywords, separated by commas ✓					
Creator * Creator of the content ✓					
Date * Date work was created/published ✓					
Collection * Community Media ✓					
Test item * No ✓					
Language * Language of the work ✓					
License * No license selected ✓					
More Options Add additional metadata					

Command line tool

The ‘command line’ tool is a new executable that can be launched from the shell of your computer. It allows you to upload, download, to do metadata operations and searches:

<http://internetarchive.readthedocs.io/en/latest/cli.html>

It is written in Python so it is usable on any OS and can be installed very easily. Once it is done, the upload will start executing, just like this:

```
$ ia upload <identifier> file1 file2 --metadata="mediatype:texts"
```

Note on OCR (optical character recognition)

For files like PDFs – resulting from a scan of text pages – IA automatically executes an optical character recognition (OCR). This is undoubtedly a feature that gives added value to this permanent digital space.

DocumentCloud

This is both a tool and a catalogue to analyse, take note, publish and file documents, that aims to transform documents in data. The URL is <https://www.documentcloud.org>.

This tool allows users to upload different formats, among which PDFs and other file types supported by LibreOffice (Microsoft Word, Excel PowerPoint, Rich Text File and other various image formats such as TIFF, PNG, GIF and JPEG).

If in the files there’s some text that could be recognized by an optical character recognition program (as PDF or images coming from a scan), DocumentCloud will try to extract it.

Each uploaded document is analysed with “Thomson Reuters OpenCalais” in order to automatically extract any names of places, persons, organizations and dates from texts.

Once uploaded, you can choose to publish them and make them be available in the public search engine. In addition, there are easy modes to embed one or more files into other web pages.

Here are two modes to upload files.

Via an online form

This is the easiest mode: a web page from which you can upload one or more files and organise them into collections (‘Project’).



Via API

DocumentCloud can be used via API¹⁰, which allows developers to integrate it within software apps and procedures.

They are very easy to use even for bulk operations involving a large number of files where the process of metadatation made by uploading could turn out to be very demanding and with a high risk of errors. Here is the documentation:

<https://www.documentcloud.org/help>

Wget

This is a free, open source command line app, very famous and widely spread: it is made to receive data and files through the most commonly used internet protocols (HTTP, HTTPS, FTP, FTPS, etc.):

<https://www.gnu.org/software/wget/>.

It has dozens of options such as those to make spider and copy operations on a website.

Site spidering allows you to extract URLs of all the pages in it to then create a copy in the archive.

For example, a sample of a command to do this with Wget is:

```
$ wget -k -K -E -r -l 10 -p -N -F -e robots=off --restrict-file-names=windows -nH http://sitoDaArchiviare.it/
```

IFTTT

This is a free service (<https://ifttt.com>) that links web apps together and/or with IOT (Internet of things) hardware to create custom actions. You can create are hundreds of them and you also have the chance to automatically file data from the web.

For example: on a spreadsheet you can archive all the elements published at the same time on an RSS feed¹¹. This way you can build a rich archive of titles, descriptions, publishing dates and URLs.

You only have to create an account and than activate the services you want to link. In this case, the RSS feed is a spreadsheet on GoogleDrive.

¹⁰https://www.wikiwand.com/it/Application_programming_interface

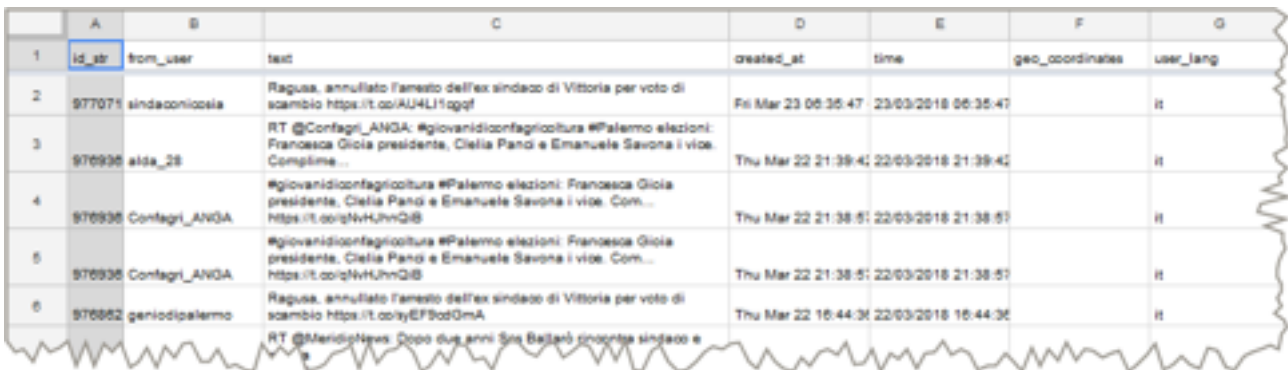
¹¹ <https://www.wikiwand.com/it/RSS>

RESOURCES AND TOOLS

TAGS

This is a free service (<https://tags.hawksey.info/>) that allows you to file on a spreadsheet all the results of a search on Twitter. It collects tweets from approximately the last 9 days; then you can make it work so it automatically updates over time, in order to collect data for long periods of time. Not all tweets are optimised or available through Twitter search interface, so TAGS does not gather the complete set of outputs.

You need a Twitter account and Google spreadsheet to use Google Sheet.



	A	B	C	D	E	F	G
1	id_str	from_user	text	created_at	time	geo_coordinates	user_lang
2	977071	sindaconiocola	Ragusa, annullato l'arresto dell'ex sindaco di Vittoria per voto di scambio https://t.co/AU4LH1ggqf	Fri Mar 23 06:35:47	23/03/2018 06:35:47		it
3	976936	aida_28	RT @Confagri_ANGA: #giovandidconfagricoltura #Palermo elezioni: Francesca Giola presidente, Clelia Pandi e Emanuele Savona i vice. Complime...	Thu Mar 22 21:39:41	22/03/2018 21:39:41		it
4	976936	Confagri_ANGA	#giovandidconfagricoltura #Palermo elezioni: Francesca Giola presidente, Clelia Pandi e Emanuele Savona i vice. Com... https://t.co/q2NtUhnQIB	Thu Mar 22 21:38:51	22/03/2018 21:38:51		it
5	976936	Confagri_ANGA	#giovandidconfagricoltura #Palermo elezioni: Francesca Giola presidente, Clelia Pandi e Emanuele Savona i vice. Com... https://t.co/q2NtUhnQIB	Thu Mar 22 21:38:51	22/03/2018 21:38:51		it
6	976862	geniodipalermo	Ragusa, annullato l'arresto dell'ex sindaco di Vittoria per voto di scambio https://t.co/yEP3odQmA	Thu Mar 22 16:44:36	22/03/2018 16:44:36		it
			RT @Meridionews: Dopo due anni Sir Balleari rincontra sindaco e				

Authors and contributors

Nicolas Kayser-Bril [@nicolaskb](#)

He is a data-driven journalist and he has been one of the first ones to practice it in Europe. He co-founded and managed [Journalism++](#) from 2011 to 2017. Before that, he was head of datajournalism at [Owni](#). He co-designed the Offshore Journalism Toolkit with Mario Tedeschini-Lalli.

Mario Tedeschini Lalli [@tedeschini](#)

He works as a digital journalist since 1997. He is a consultant on Digital Editorial strategies, visiting lecturer in Digital Journalism at the Urbino Journalism School and founder of the Italian group of the [Online News Association](#). He is Nicolas Kayser-Bril partner of the Offshore Journalism Toolkit project.

Dragan Espenschied [@despens](#)

He is a media artist, home computer folk musician, digital culture researcher and conservator. Since April 2014, he is leading the Digital Conservation Program at the internet arts organization [Rhizome](#), introducing a practice-based approach to preserving the institution's collection and the ingest of new artifacts.

Guy Berger [@guyberger](#)

He is UNESCO's director for [Freedom of Expression and Media Development](#). Previously he headed the School of Journalism and Media Studies at Rhodes University, South Africa. He has also worked in both press and television and had a long-running column on the The Mail & Guardian website.

William Kilbride [@WilliamKilbride](#)

He is Executive Director of the [Digital Preservation Coalition](#). He started his career as an archaeologist in the early 1990s when the discipline's enthusiasm to use technology was not matched with the skills to look after the resulting data. This gave him an early and practical introduction to the challenges of data management and preservation.

AUTHORS AND CONTRIBUTORS

Claudio Agosti [@_vecna](#)

He is a software developer with a focus on explaining and researching into the abstractness of algorithm surveillance and data marketing. He is a member of the [Good Technology Collective](#) and the Berlin-based think tank [Diem25](#). He contributed to the documentary [Nothing to Hide](#) and is vice president of the [Hermes Center](#).

Ernesto Belisario [@diritto2punto0](#)

He is a lawyer of the [E-Lex](#) law firm and he is specialised in new technology law - data protection and privacy, open government, start-up, social media - and in administrative law. He is member of the Italian Government's Board on "Innovation and Digital Agenda".

Andrea Borruso [@aborruso](#)

He is a geospatial analyst. He is the president of [onData](#), a nonprofit organization focused on open-data, civic technology and investigative journalism. He was in charge of software development for Panoptes, a company that built drones. He is among the authors of the open data guidelines of the Town Councils of Palermo and Matera. In his free time he is a civic hacker.

www.offshorejournalism.com



This report is published under a Creative Commons licence (CC BY-SA 4.0)