Archiving Social Media

Talking about open data and social archives

Aniol Maria Vallès, member of the Working Group on Data and Transparency (Society of Catalan Archivists and Records Managers)
Participate with Slido

https://www.sli.do
Event code: #ArchivoSocial
5 ★ Open Data

Tim Berners-Lee (http://5stardata.info):

- ★ make your stuff available on the Web (whatever format) under an open license
- ★★ make it available as structured data (Excel instead of a scanned image)
- ★★★ use non-proprietary formats (CSV and ODF instead of Excel)
- ★★★★ use URIs to denote things, so that people can point at your stuff
- ★★★★★ link your data to other data to provide context
Social fonds

Understanding the Social fonds

«the entire body of records created by a community of users through a participatory dynamic in social platforms around a shared interest or event»

Walking the path between archival activism and the social archive(s) Vicenç Ruiz Society of Catalan Archivists and Records Managers
Social fonds

Understanding the Social fonds

«Are produced by a collective creator (their only bond may sometimes be as platform users, in a private system not owned by any single creators (digital colonialism); it is massive and brief»

Walking the path between archival activism and the social archive(s)
Vicenç Ruiz Society of Catalan Archivists and Records Managers
Social fonds

Understanding the Social fonds

«Is a priority to design a participatory recordkeeping model, community archives (Monash PRM), to act proactively so as to guarantee a democratic and free long-term access (activist data archiving), and to develop tools that allow for automatization of the analysis processes (appraisal, description,...) and for exploitation of information (Computational Archival Science)»

Walking the path between archival activism and the social archive(s)
Vicenç Ruiz Society of Catalan Archivists and Records Managers
Social websites

Centralized social websites

"Social Media Illustrations" by Isaac Anthony is licensed under CC BY-NC-ND 4.0
https://search.creativecommons.org/photos/e9571c58-0453-4df6-a802-3f67dda579bc
Datasets

Datasets are packages with data sets selected from a Big Data file.
Web formats

CSV
«Document in simple open format to represent data in table form, in which the columns are separated by commas and the rows by line breaks»

JSON
«Open and standard text-based language, designed for human-readable data exchange, derived from JavaScript scripting language, and enabling the representation of simple data structures and associative lists, called objects»
**Metadata** is all the descriptive information about the context, quality, condition or characteristics of a resource, data or object, in order to facilitate its recovery, authentication, evaluation, preservation and / or interoperability.

https://commons.wikimedia.org/wiki/File:Donn%C3%A9es_Exif.jpg
Archive activism

Witness https://www.witness.org/resources/
Website archives

ArchiveBox takes a list of website URLs you want to archive, and creates a local, static, browsable HTML clone of the content from those websites (it saves HTML, JS, media files, PDFs, images and more).

ArchiveBox https://archivebox.io
Tweet JSON

We want to hear from you! The @ICAriChiv calls on us to reflect together on the role and social function of the #archives. Use #AnArchivels to exchange ideas and diversify the discussion. What is an archive is for you...?

#AnArchivels #UneArchiveCest #UnArchivoEs - David Fricker (I... What is an Archive? What does it mean to the records and archives communities? What does it mean to our users? How... youtube.com

11:30 a.m. · 24 d’abr. de 2020 · Hootsuite Inc.

17 Retuits 25 Agradaments

https://twitter.com/ICAriChiv/status/1253617253351817216

www.ica.org #IAW2020
Tweet JSON

https://twitter.com/ICArchiv/status/125361725335181721

6

www.ica.org  #IAW2020
Because we don’t have a data transfer contract, we have to rely on what the API (Application Programming Interface) reflects and the sample it offers in JSON.
Use of the API

API (Application Programming Interface)

- An interface that specifies how different components of computer programs should interact.

- A set of indications, in terms of functions and procedures, offered by a service to be used by another program to interact with the program in question.

- A set of statements that defines the contract of a computer component with whom it will use its services.
Use of the API

«An API is a set of indications, functions and procedures that allow components of computer programs to interact.»
## Metadata

### Comparison Twitter metadata with various schemes

<table>
<thead>
<tr>
<th>EAD</th>
<th>Twitter</th>
<th>Dublin Core</th>
<th>PREMIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;accesrestrict&gt;</td>
<td>withheld_copyright</td>
<td>rights</td>
<td>rightsStatement</td>
</tr>
<tr>
<td>&lt;author&gt;</td>
<td>user</td>
<td>creator</td>
<td>linkingAgentIdentifier</td>
</tr>
<tr>
<td>&lt;date&gt;</td>
<td>created_at</td>
<td>date</td>
<td>eventDateTime</td>
</tr>
<tr>
<td>&lt;language&gt;</td>
<td>lang</td>
<td>language</td>
<td>-</td>
</tr>
<tr>
<td>&lt;unitid&gt;</td>
<td>id</td>
<td>identifier</td>
<td>objectIdentifier</td>
</tr>
</tbody>
</table>
Dehydrate & Hydrate

1253617253351817216
How unique tweet identifiers are generated

1164106327217639425

Convert the decimal id to binary

1000000100111101111001000100101110010100101111001000000000001

Cómo se generan los identificadores únicos de los tweets - El Blog de Marcelino Madrigal
https://www.mmadrigal.com/como-se-generan-los-identificadores-unicos-de-los-tweets/
How unique tweet identifiers are generated

100000010011110111100100100101110010100101111001000000000001

Data Center where the tweet was generated:
01011 = Data Center Nº 11

Server where was generated the tweet:
11001 = Servidor 25

Ordinal number of the tweet

This ID code would be the first tweet from server 25 of Data Center 11

Cómo se generan los identificadores únicos de los tweets - El Blog de Marcelino Madrigal
https://www.mmadrigal.com/como-se-generan-los-identificadores-unicos-de-los-tweets/
How unique tweet identifiers are generated

100000010011110111100100101110010101011110010000000000001

We convert the binary number to decimal

277544576458 + 1288834974657 = 1566379551115

(a Twitter check “hash” number by Snowflake)

Conversion to format Epoch

GMT: Wednesday, 21 August 2019 9:25:51.115

Cómo se generan los identificadores únicos de los tweets - El Blog de Marcelino Madrigal
https://www.mmadrigal.com/como-se-generan-los-identificadores-unicos-de-los-tweets/
Visualization

#UniteTheRight

Gephi dataviz of #UniteTheRight deleted tweets as a result of Twitter suspending accounts
Visualization

http://tweettopicexplorer.neoformix.com/#n=ICA#Archiv
Visualization

http://www.proyectocuentalo.org
Archival projects with Twitter

Library of Congress example

- On April 14, 2010, the Library of Congress decided to preserve all public tweets that have been created on Twitter over last years and that are worthy of preservation.

- The donation will be through the Gnip company that would transfer the tweets to the Library of Congress storage systems.

- There were more than 55 million daily messages, 50 metadata fields,... Of “free” access, indexing, reliability, validity and usefulness.
Archival projects with Twitter

Library of Congress example

- The total collection of tweets includes the first 12 years of the social network (2006-2017)
- On December 26, 2017, it was decided to be more selective when capturing and collecting tweets.
- The Library of Congress collection does not include images, videos or linked content
- Access to the collection remains embargoed until it is decided how this access will be

Archival projects with Twitter

UK National Archives example

- Interesting but not an archive project like the Library of Congress right now. It is more a purposeful exhibition and collection.

- They offer a view of the "live" tweets in JSON, XML with a direct link to the official Twitter accounts.

- It also promises captured links within tweets, though not linked content. The link will be shown indicating where the resource simply was.
Archival projects with Twitter

The UK National Archives example
Archival projects with Twitter

Social Control 2019 Zine
A guide to owning your social media content - Alexandra Dolan-Mescal


Download your Facebook Archive and donate it to the Alexander Turnbull Library

https://www.youtube.com/watch?v=-2FHw23Q4SQ

https://twitter.com/fromADMwithlove/status/1156960213033017344
Archival projects with Twitter

Peticiones con el Reglamento General de Protección de Datos (GDPR)/Petition about the General Data Protection Regulation (GDPR)

- Facebook scraped calls, text message data for years from Android phones [Updated] | Ars Technica

- How to Find Out Everything Facebook Knows About You

Dylan McKay al Twitter: "Downloaded my facebook data as a ZIP file Somehow it has my entire call history with my partner's mum
https://t.co/CIRUguf4vD" / Twitter
https://twitter.com/dylanmckaynz/status/976368845635035138/

Number: 61234

<table>
<thead>
<tr>
<th>Call Type</th>
<th>Start time</th>
<th>Duration</th>
<th>Name</th>
<th>Number Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUTGOING</td>
<td>Monday, 26 Nov 2016 at 21:57 UTC+13</td>
<td>2</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>OUTGOING</td>
<td>Sunday, 10 Apr 2017 at 10:53 UTC+12</td>
<td>42</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>MISSED</td>
<td>Monday, 10 Feb 2017 at 18:18 UTC+13</td>
<td>0</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>OUTGOING</td>
<td>Tuesday, 29 Nov 2016 at 17:38 UTC+13</td>
<td>446</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>INCOMING</td>
<td>Sunday, 26 Mar 2017 at 11:23 UTC+13</td>
<td>120</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>OUTGOING</td>
<td>Sunday, 26 Mar 2017 at 12:37 UTC+13</td>
<td>115</td>
<td>Mereana Gell</td>
<td></td>
</tr>
<tr>
<td>INCOMING</td>
<td>Saturday, 6 May 2017 at 9:39 UTC+13</td>
<td>39</td>
<td>Mereana Gell</td>
<td></td>
</tr>
</tbody>
</table>
Twarc

Collect Twitter Data with Twarc! · Learn Twarc!
https://scholarslab.github.io/learn-twarc/
Publish datasets

Tweet ID Datasets http://www.docnow.io/catalog/

Twitter's terms of service don't allow tweet datasets to be published on the web, but they do allow tweet identifier datasets to be shared. This speaks to users rights as content creators, while also allowing researchers to share their data with others.

This site is a catalog of datasets that are publicly available on the web. If you would like to turn these tweet identifier datasets back into the original JSON first download the dataset and then use the Hydrator desktop application, or Twarc if you are comfortable working at the command line.

You can add your own datasets to the catalog by following these instructions. If you'd like updates when datasets are added please subscribe to the RSS feed. All metadata listed here is licensed CC0. You may want to refer to our code of conduct if you have questions or concerns about the datasets we list here.
Publish datasets

TweetSets - The George Washington University Dataverse https://tweetsets.library.gwu.edu

Other datasets that we’ve collected:
- Trump Administration officials
- Healthcare
- Make America Great Again
- Tweets to @realDonaldTrump
- Tax reform
- Trump inauguration
- A Day Without A Woman
- Alts and rogues
- Immigration and travel ban executive order
- March for Science
- Charlottesville
- Solar Eclipse

Contact me if you’d like any of these loaded into TweetSets.

Basat en: Social Feed Manager
https://gwu-libraries.github.io/sfm-ui/
Publish datasets

Social Feed Manager https://gwu-libraries.github.io/sfm-ui/
Conclusions

- A tweet is not the image that comes to us from the browser but the entire set of metadata that Twitter is hosting.

- We can know what Twitter is hosting and how it could be archived thanks to the API and JSON.

- We can validate Twitter messages if we receive enough metadata from the tweet or Twitter profile.
Audience Q&A Session

ℹ️ Start presenting to display the audience questions on this slide.