IPTC Photo Metadata Crawler

Brendan Quinn, Managing Director, IPTC

IPTC Photo Metadata Conference 2019

Paris 6 June 2019
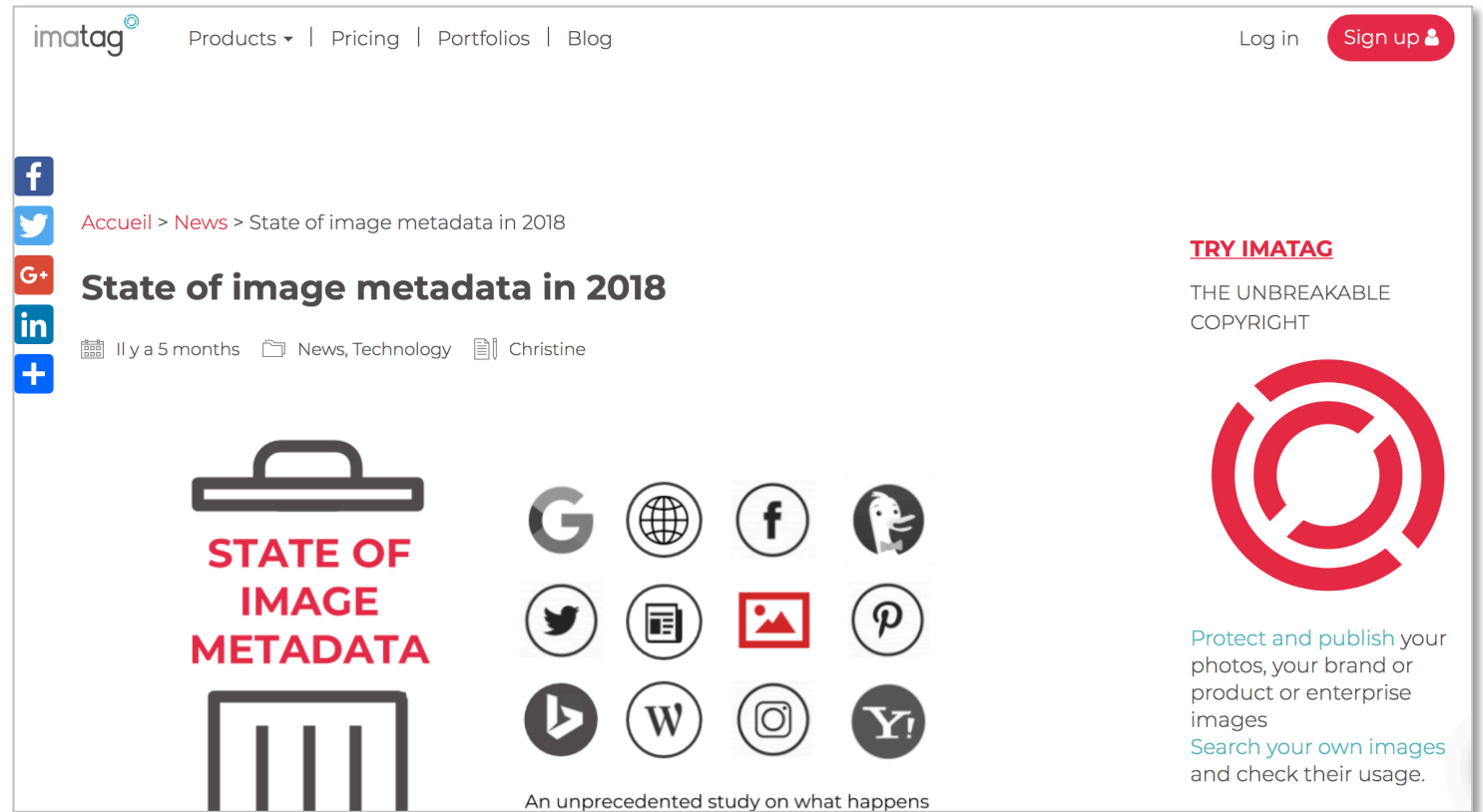
The problem

- We know that some sites are better at implementing IPTC Photo Metadata than others, but…

- Who can we point to as examples of best practice?

- Who should we remind to improve their metadata? (particularly in the news and media space where we are focussed)

- How can we compare sites' usage of IPTC Photo Metadata?
  - who uses which fields?
  - who is putting what text in their fields?

- How can we track sites' usage of IPTC Photo Metadata over time?
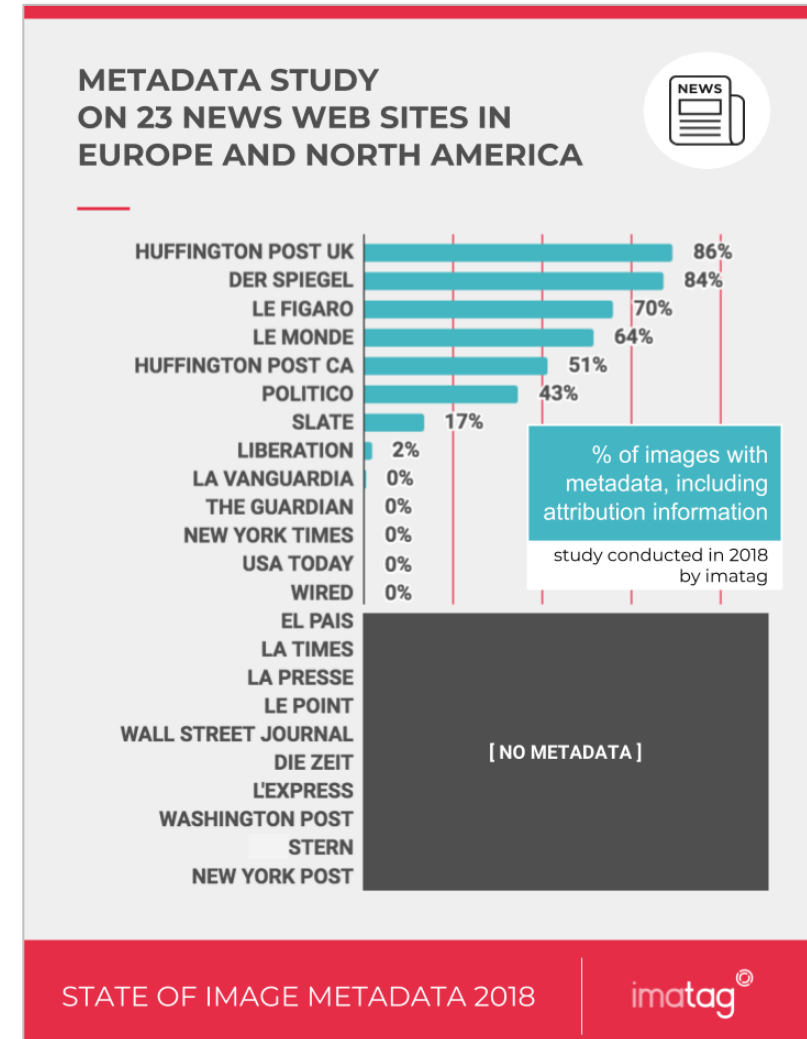
The inspiration

- At the IPTC Photo Metadata Conference in May 2018, Imatag presented their survey of embedded image metadata
- Their crawler identified that 85% of images published on the web have no metadata, and only 3% have copyright information
  - Based on "a sample of over 40 million images, not including social media or image databases"
- The figures become 80% and 8% on "editorial websites"
  - "More than 50,000 images have been analyzed by IMATAG on over 750 different editorial websites worldwide."

imatag   Products ▾ | Pricing | Portfolios | Blog   Log in   Sign up

Accueil > News > State of image metadata in 2018

State of image metadata in 2018

Il y a 5 months   News, Technology   Christine

STATE OF IMAGE METADATA

An unprecedented study on what happens

TRY IMATAG
THE UNBREAKABLE COPYRIGHT

Protect and publish your photos, your brand or product or enterprise images
Search your own images and check their usage.

Charts with specific news providers are very useful

- Showing usage of metadata across news providers:
  - Those that use metadata well, including description and rights information
  - Those that use other metadata (EXIF)
  - Those that use no metadata at all
- This might be for a variety of reasons…
  - Not receiving images with metadata from agencies or photographers
  - Deliberate stripping of metadata for privacy reasons, or to protect sources
  - CMS, DAM, or CDN tools may be stripping out metadata inadvertently or deliberately

METADATA STUDY ON 23 NEWS WEB SITES IN EUROPE AND NORTH AMERICA

News Site	%
HUFFINGTON POST UK	86%
DER SPIEGEL	84%
LE FIGARO	70%
LE MONDE	64%
HUFFINGTON POST CA	51%
POLITICO	43%
SLATE	17%
LIBERATION	2%
LA VANGUARDIA	0%
THE GUARDIAN	0%
NEW YORK TIMES	0%
USA TODAY	0%
WIRED	0%
EL PAIS	[NO METADATA]
LA TIMES	
LA PRESSE	
LE POINT	
WALL STREET JOURNAL	
DIE ZEIT	
L'EXPRESS	
WASHINGTON POST	
STERN	
NEW YORK POST	

% of images with metadata, including attribution information

study conducted in 2018 by imatag

STATE OF IMAGE METADATA 2018 | imatag

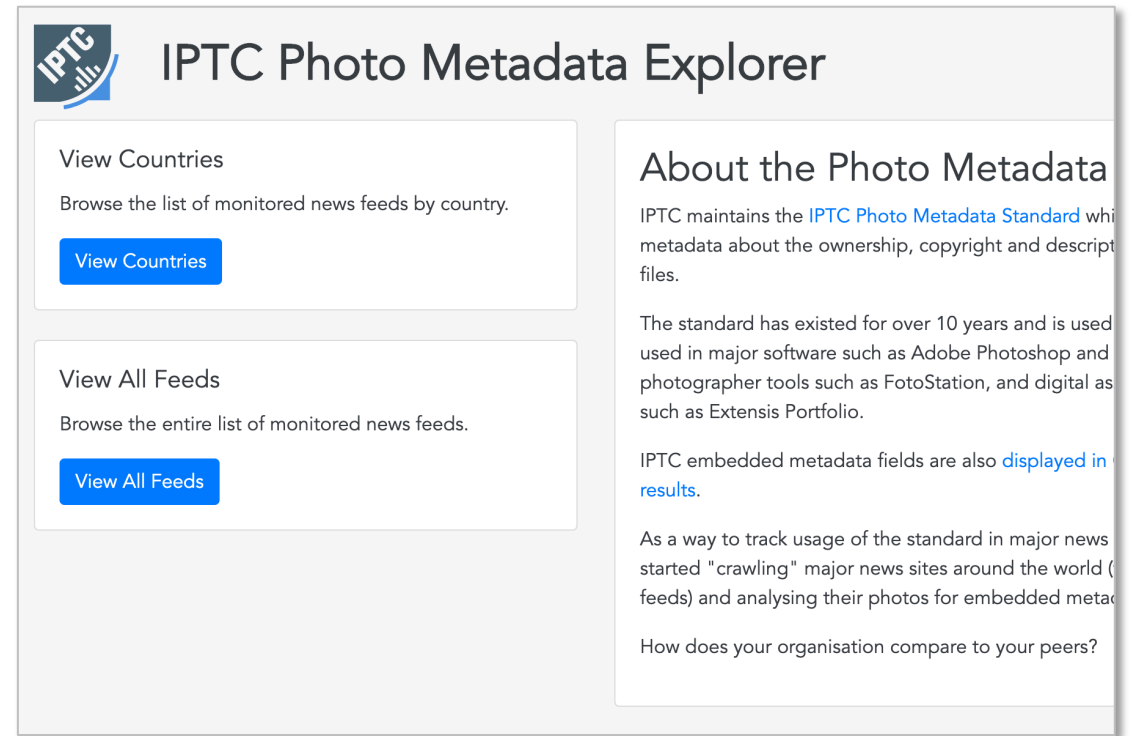# IMATAG's search can also show embedded metadata
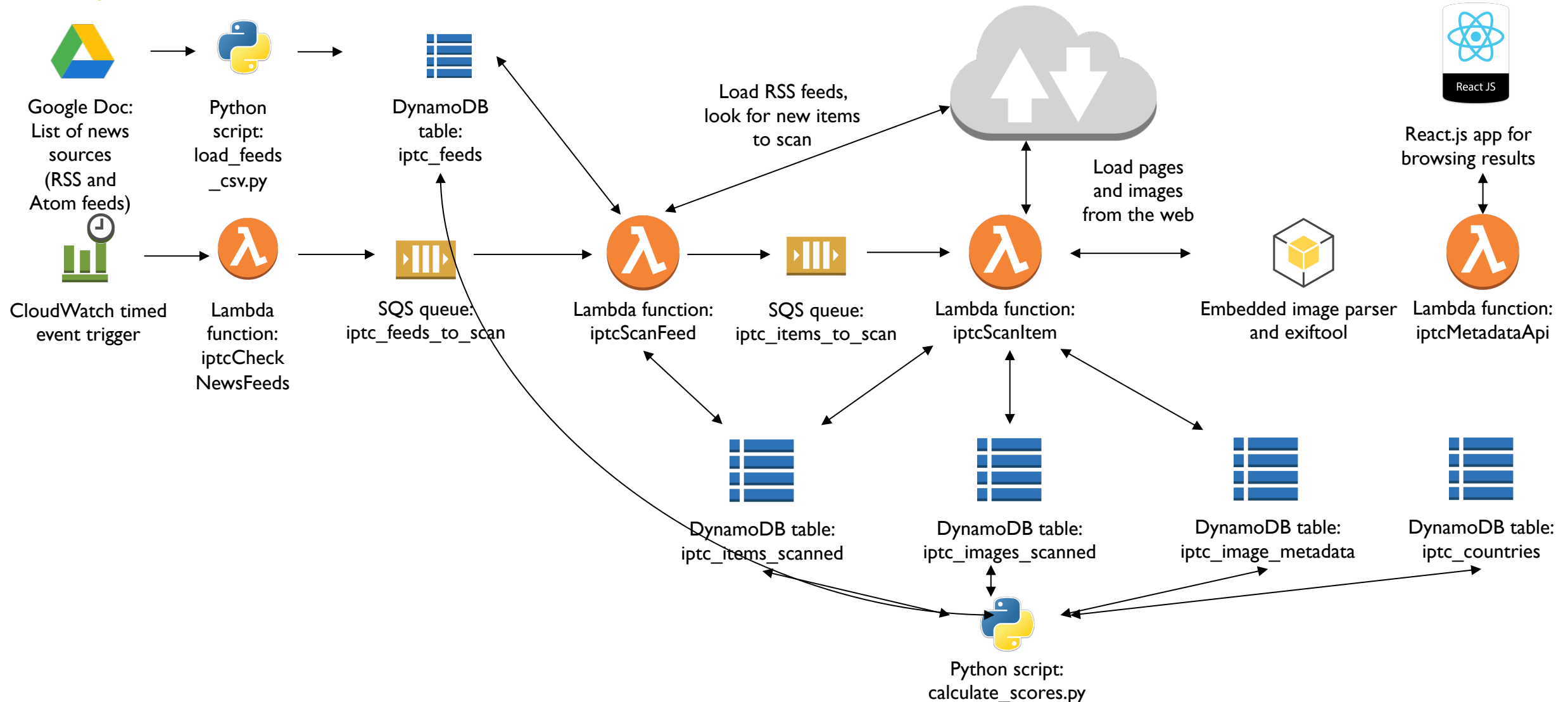

- http://www.imatag.com/search/

5

This was inspiration for our own crawler

- We have the photo metadata checker tool (pictured) but that only works on one image or one page at a time
- We thought it would be useful to have something that tracks IPTC Photo Metadata usage on news sites
- So why not build something simple and lightweight that can track image metadata usage?
- We can run it as often as we like to track trends over time
- Lets us analyse usage, find best practices and people who could be encouraged to follow the spec better
- Based on news publishers' public RSS or Atom feeds

IPTC Photo Metadata Explorer

View Countries

Browse the list of monitored news feeds by country.

[View Countries]

View All Feeds

Browse the entire list of monitored news feeds.

[View All Feeds]

About the Photo Metadata

IPTC maintains the IPTC Photo Metadata Standard whi metadata about the ownership, copyright and descript files.

The standard has existed for over 10 years and is used used in major software such as Adobe Photoshop and photographer tools such as FotoStation, and digital as such as Extensis Portfolio.

IPTC embedded metadata fields are also displayed in results.

As a way to track usage of the standard in major news started "crawling" major news sites around the world (feeds) and analysing their photos for embedded meta

How does your organisation compare to your peers?

IPTC Photo Metadata Crawler Architecture

Google Doc: List of news sources (RSS and Atom feeds)

Python script: load_feeds_csv.py

DynamoDB table: iptc_feeds

Load RSS feeds, look for new items to scan

Load pages and images from the web

React.js app for browsing results

CloudWatch timed event trigger

Lambda function: iptcCheckNewsFeeds

SQS queue: iptc_feeds_to_scan

Lambda function: iptcScanFeed

SQS queue: iptc_items_to_scan

Lambda function: iptcScanItem

Embedded image parser and exiftool

Lambda function: iptcMetadataApi

DynamoDB table: iptc_items_scanned

DynamoDB table: iptc_images_scanned

DynamoDB table: iptc_image_metadata

DynamoDB table: iptc_countries

Python script: calculate_scores.py

Checking 200 news feeds

Country	Number of feeds checked	Examples
AT	8	DerStandard, ORF
CA	4	CBC, CTV, Radio Canada, 500px…
CH	2	Neue Zürcher Zeitung, Blick
DE	18	Der Spiegel (ENG/DE), Bild, DW, Die Zeit…
ES	6	El Pais, El Mundo, ABC…
FR	8	Le Monde, Le Figaro, France24…
GB	21	BBC, Times, Daily Mail, Metro, Economist…
IN	5	Times of India, Hindustan Times, The Hindu…
IT	6	ANSA, RAI News
JP	17	Japan Times, Nikkei Shinbun, Yomiuri Shinbun
US	9	CNN, NYT, USA Today, observer.com

Every EU country has at least three feeds

A RESTful API to the data
lets us build an interactive Photo Metadata Explorer tool

Feeds by country (sortable)

IPTC Photo Metadata Explorer

Feeds in DE

Feed name	country	Scanned date	Item Count	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing
taz	DE	24 May 2019	100	87.9%	75.4%	87.9%
Spiegel Online (German)	DE	24 May 2019	100	82.1%	53.7%	80.0%
Spiegel Online International	DE	24 May 2019	86	56.6%	21.2%	50.6%
The Local	DE	24 May 2019	100	11.4%	11.3%	11.3%
Munich NOW	DE	24 May 2019	13	3.1%	1.7%	3.1%
Frankfurter Allgemeine	DE	24 May 2019	100	1.5%	1.0%	1.5%
tagesschau.de	DE	24 May 2019	100	1.0%	0.8%	1.0%
Frankfurter Rundschau	DE	24 May 2019	0	0.0%	0.0%	0.0%
Bild	DE	24 May 2019	100	0.0%	0.0%	0.0%
stern.de	DE	24 May 2019	100	0.0%	0.0%	0.0%
ZDF Heute	DE	24 May 2019	95	0.0%	0.0%	0.0%

Browsing to individual images

IPTC Photo Metadata Explorer

Feed: taz

(Feed percentage: 87.9%)

Articles in this feed:

Article title	Scanned date	Image count	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
Finanzen der extremen Rechten: Die Anti-Europäer	2019-05-09	1	100.0%	0.0%	100.0%	100.0%
Sudan nach der Revolution: Mit Geduld in die Freiheit	2019-05-09	1	100.0%	100.0%	100.0%	100.0%
Atomabkommen mit Iran: Teheran platzt der Kragen	2019-05-09	2	50.0%	50.0%	50.0%	50.0%
Die Westdeutschen und die Systemfrage: Kampfbegriff Sozialismus	2019-05-09	1	100.0%	100.0%	100.0%	100.0%
Datenschutz bei Gentests: Gefährliches Wissen	2019-05-	1	100.0%	100.0%	100.0%	100.0%

Browsing to individual images

IPTC Photo Metadata Explorer

Feed: taz
(Feed percentage: 87.9%
Articles in this feed:

Article title
Finanzen der extremen Rechten: Die Anti-
Sudan nach der Revolution: Mit Geduld in
Atomabkommen mit Iran: Teheran platzt d
Die Westdeutschen und die Systemfrage: Sozialismus
Datenschutz bei Gentests: Gefährliches W

IPTC Photo Metadata Explorer
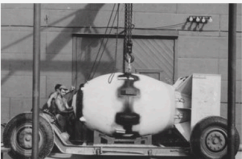
Article: Atomabkommen mit Iran: Teheran platzt der Kragen

Scanned on 9 May 2019

Original URL: http://www.taz.de/Atomabkommen-mit-Iran/!5589921/

List of images in this article:

Image (click for details)	Scanned date	IPTC Descriptive Count	IPTC Things Shown Count	IPTC Rights & Licensing	IPTC Metadata?
	2019-05-09	10	6	4	true
	2019-05-09	0	0	0	false

Browsing to individual images

IPTC Photo Metadata is split into groups:
Descriptive, Things Shown, Rights & Licensing

Limitations and exceptions

- Only works for sites with public RSS or Atom feeds
- Won't work if the main image is smaller than 150x150 pixels
- Won't work on embedded images (e.g. posts with only Twitter, Instagram or YouTube embeds)
- Editorial visualisations and infographics rarely have metadata attached (but they should!)
- More and more articles use video rather than images which won't be scanned
- Won't re-read metadata if it is changed after we have scanned a page
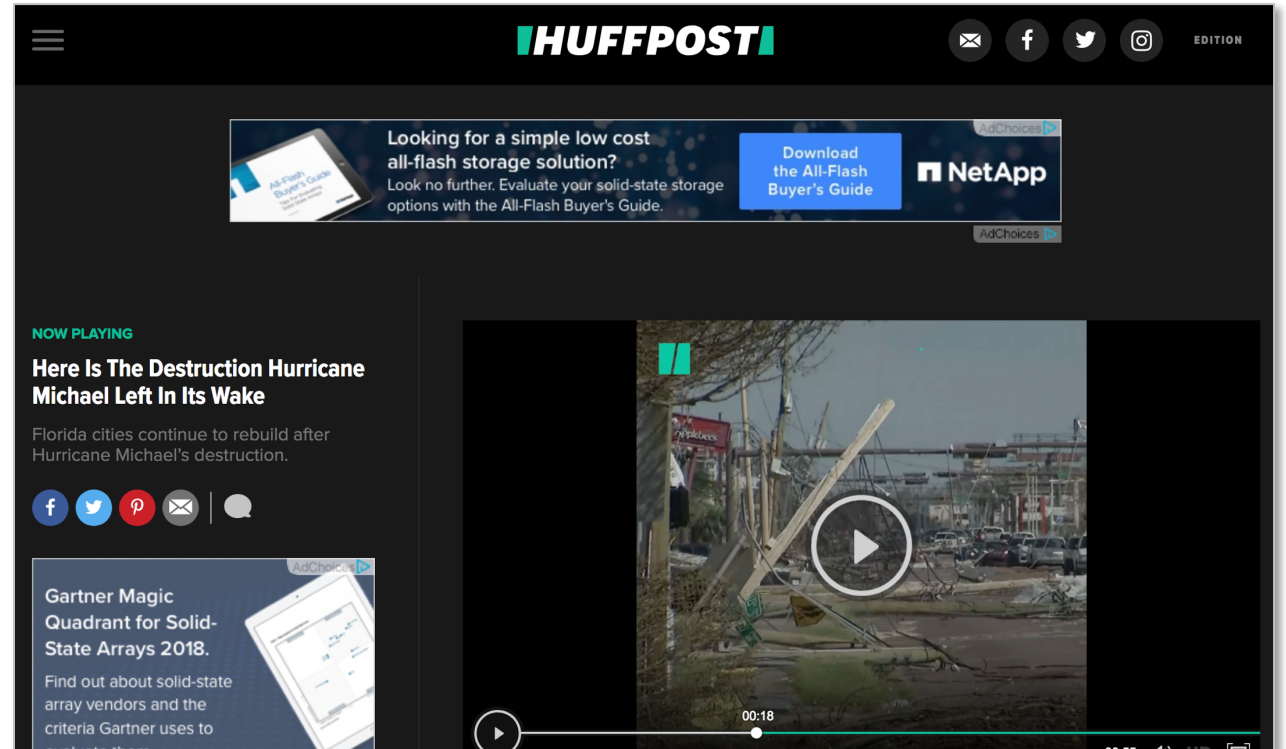
How do we calculate scores?

At image level, we count occurrences of metadata tags form both XMP and IIM embedded metadata in 3 categories:

Descriptive	Things shown	Rights & LIcensing
• Title (XMP IPTC-Core 6.16), Object Name (IIM 2:05) • Headline (XMP IPTC-Core 6.5 / IIM 2:105) • Description (XMP IPTC-Core 6.4) • Keywords (XMP IPTC-Core 6.7 / IIM 2:25) • Caption/Abstract (IIM 2:120) • Date Created (XMP IPTC-Core 6.12 / IIM 2:55) • Time Created (IIM 2:60) • Category (IIM 2:15) • Supplemental Categories (IIM 2:20) • **Creator (IIM 2:80 By-line),** • Byline Title (IIM 2:85)	• Location (XMP and IIM Location code, city/province/state • Person (XMP PersonInImage, PersonInImageDescription, PersonInImageCVTerm) • Organisation (XMP OrganisationInImageCode, OrganisationInImageName) • Artwork (XMP ArtworkOrObject, Artwork Circa Date Created, ArtworkCreator,, ArtworkStylePeriod, ArtworkTitle, …) • Product (ProductInImage, ProductInImageGTIN)	• **Credit (IIM 2:110) / Credit Line (XMP IPTC-Core 6.21)** • Source (XMP IPTC-Core 6.23) • **Copyright Notice (IIM 2:116)** • CopyrightYear • ArtworkCopyrightNotice, ArtworkCopyrightOwnerID, ArtworkCopyrightOwnerName • Model release (Model Age, Model Additional Information) • Artwork Licensor • Embedded Encoded Rights Expression / Embedded Linked Rights Expression • Registry ID / Registry Item ID

Calculating "scores" for pages, feeds and countries

- Individual images:
  - Simply count all occurrences in each category

Calculating "scores" for pages, feeds and countries

- Individual images:
  - Simply count all occurrences in each category

- Pages ("Items"):
  - The percentage of images containing *any* metadata in that category

IPTC Photo Metadata Explorer

Article: L'empereur Naruhito du Japon accède officiellement au trône

Scanned on 1 May 2019

Original URL: http://www.lefigaro.fr/international/l-empereur-naruhito-du-japon-accede-officiellement-au-trone-20190501

List of images in this article:

Image (click for details)	Scanned date	IPTC Descriptive Count	IPTC Things Shown Count	IPTC Rights & Licensing	IPTC Metadata?
	2019-05-01	9	3	1	true
	2019-05-01	8	3	1	true

Calculating "scores" for pages, feeds and countries

- Individual images:
  - Simply count all occurrences in each category

- Pages ("Items"):
  - The percentage of images containing *any* metadata in that category
  - "IPTC Metadata" score is the percentage of images that contain any IPTC metadata fields at all

IPTC Photo Metadata Explorer

Home / Feeds / Le Figaro

Feed: Le Figaro
(Feed percentage: 72.1%)
Articles in this feed:

Article title	Scanned date	Image count	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
La Corée du Nord s'impatiente et procède à deux tirs de missile	2019-05-09	1	100.0%	100.0%	100.0%	100.0%
Le «chirurgien de l'horreur» soupçonné pour la première fois du décès d'un patient	2019-05-09	0	0.0%	0.0%	0.0%	0.0%
Chine: comment le Parti communiste entend désengorger les cimetières	2019-05-09	1	100.0%	100.0%	100.0%	100.0%
«Unis, nous sommes plus forts»: les 27 s'engagent pour l'avenir de l'Union européenne	2019-05-09	0	0.0%	0.0%	0.0%	0.0%
Entre les fonctionnaires et Macron, le désamour ne date pas de mai 2017	2019-05-09	1	100.0%	100.0%	100.0%	100.0%

Calculating "scores" for pages, feeds and countries

- Individual images:
  - Simply count all occurrences in each category

- Pages ("Items"):
  - The percentage of images containing *any* metadata in that category
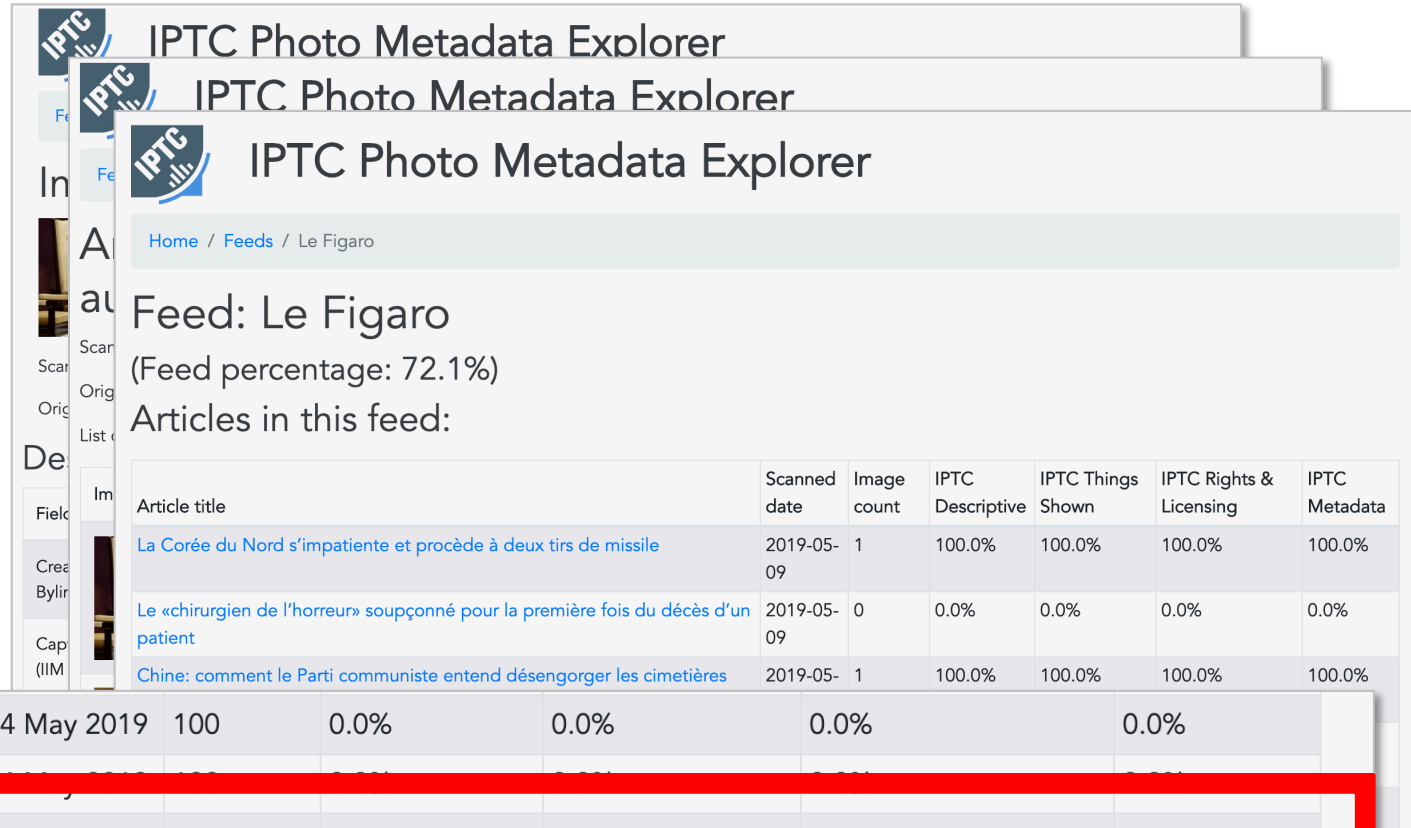 - "IPTC Metadata" score is the percentage of images that contain any IPTC metadata fields at all

- Feeds:
  - Average of percentages on all items in that feed (ignoring items with no images)

IPTC Photo Metadata Explorer

IPTC Photo Metadata Explorer

IPTC Photo Metadata Explorer

Home / Feeds / Le Figaro

Feed: Le Figaro
(Feed percentage: 72.1%)
Articles in this feed:

Article title	Scanned date	Image count	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
La Corée du Nord s'impatiente et procède à deux tirs de missile	2019-05-09	1	100.0%	100.0%	100.0%	100.0%
Le «chirurgien de l'horreur» soupçonné pour la première fois du décès d'un patient	2019-05-09	0	0.0%	0.0%	0.0%	0.0%
Chine: comment le Parti communiste entend désengorger les cimetières	2019-05-	1	100.0%	100.0%	100.0%	100.0%

La Stampa	IT	24 May 2019	100	0.0%	0.0%	0.0%	0.0%
Le Figaro	FR	24 May 2019	100	70.1%	63.1%	62.1%	72.1%
Le Monde	FR	24 May 2019	100	3.8%	3.3%	3.7%	3.8%
Le Quotidien	LU	24 May 2019	100	24.1%	17.5%	24.0%	24.1%
Le Soir	BE	24 May 2019	100	0.0%	0.0%	0.0%	0.0%

Some comparisons: IPTC metadata usage by country

IPTC Photo Metadata Explorer

All Countries

Country name	Country code	Number of feeds	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
Slovenia	SI	3	22.5%	20.5%	21.2%	22.5%
Hungary	HU	4	19.7%	9.9%	13.0%	19.8%
France	FR	6	12.3%	11.1%	13.8%	15.4%
Germany	DE	18	13.5%	9.2%	13.1%	13.7%
United States	US	9	11.6%	6.3%	9.9%	11.8%
Luxembourg	LU	3	11.3%	5.9%	8.0%	11.3%
Greece	GR	3	10.6%	10.3%	10.6%	10.6%
Italy	IT	6	8.6%	6.3%	8.4%	9.0%
Canada	CA	4	4.7%	0.1%	3.8%	4.7%
Cyprus	CY	3	4.4%	2.8%	3.4%	4.4%
Australia	AU	9	4.2%	1.9%	3.9%	4.2%
Latvia	LV	3	2.0%	1.4%	1.6%	2.0%

IPTC Photo Metadata usage by country
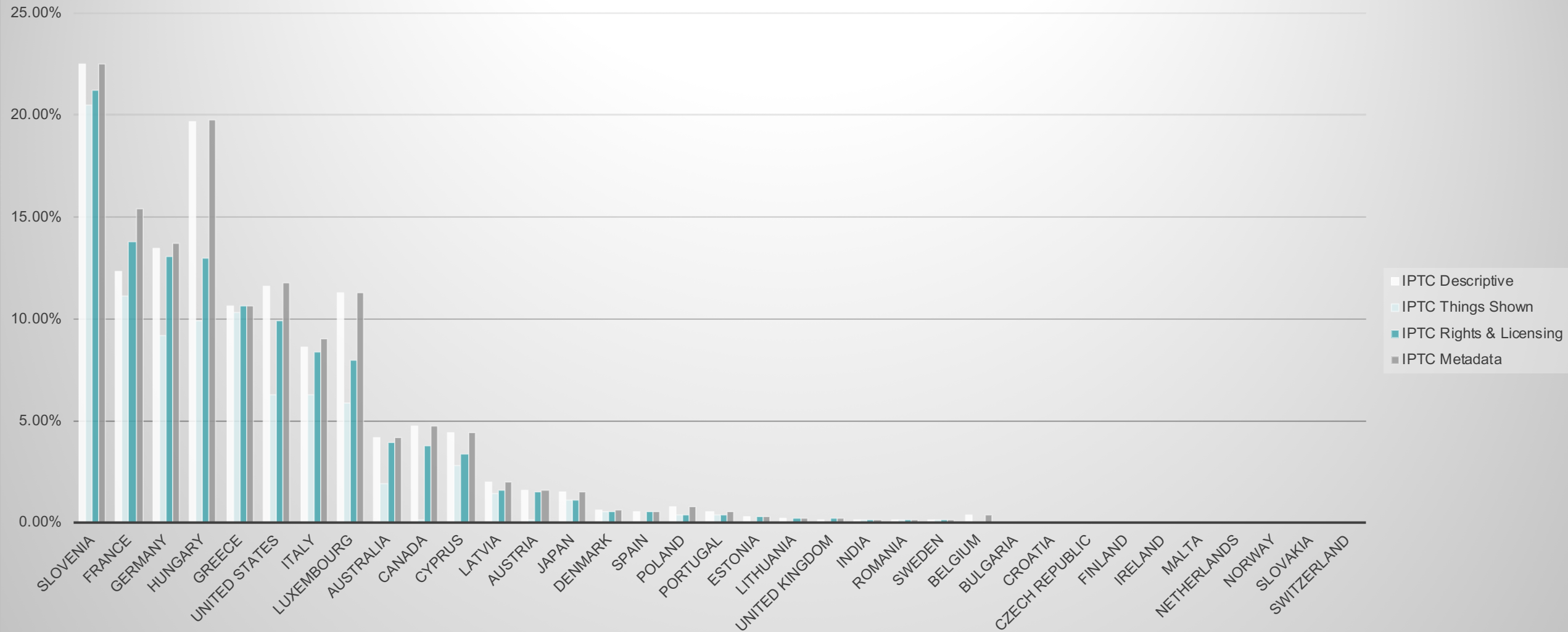
IPTC Photo Metadata - Top 10 countries

Rank	Country name	Number of feeds	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
1	Slovenia	3	22.50%	20.50%	21.20%	22.50%
2	Hungary	4	19.70%	9.90%	13.00%	19.80%
3	France	6	12.30%	11.10%	13.80%	15.40%
4	Germany	18	13.50%	9.20%	13.10%	13.70%
5	United States	9	11.60%	6.30%	9.90%	11.80%
6	Luxembourg	3	11.30%	5.90%	8.00%	11.30%
7	Greece	3	10.60%	10.30%	10.60%	10.60%
8	Italy	6	8.60%	6.30%	8.40%	9.00%
9	Canada	4	4.70%	0.10%	3.80%	4.70%
10	Cyprus	3	4.40%	2.80%	3.40%	4.40%

Notes:

- Slovenia's and Hungrary's high scores are skewed by the fact that one of the few feeds studied has high metadata usage (can be fixed by analysing more feeds)
- France and Germany have relatively high metadata usage, especially given how many feeds we're analysing
- Number 10 in the list has

Lots of room for improvement…

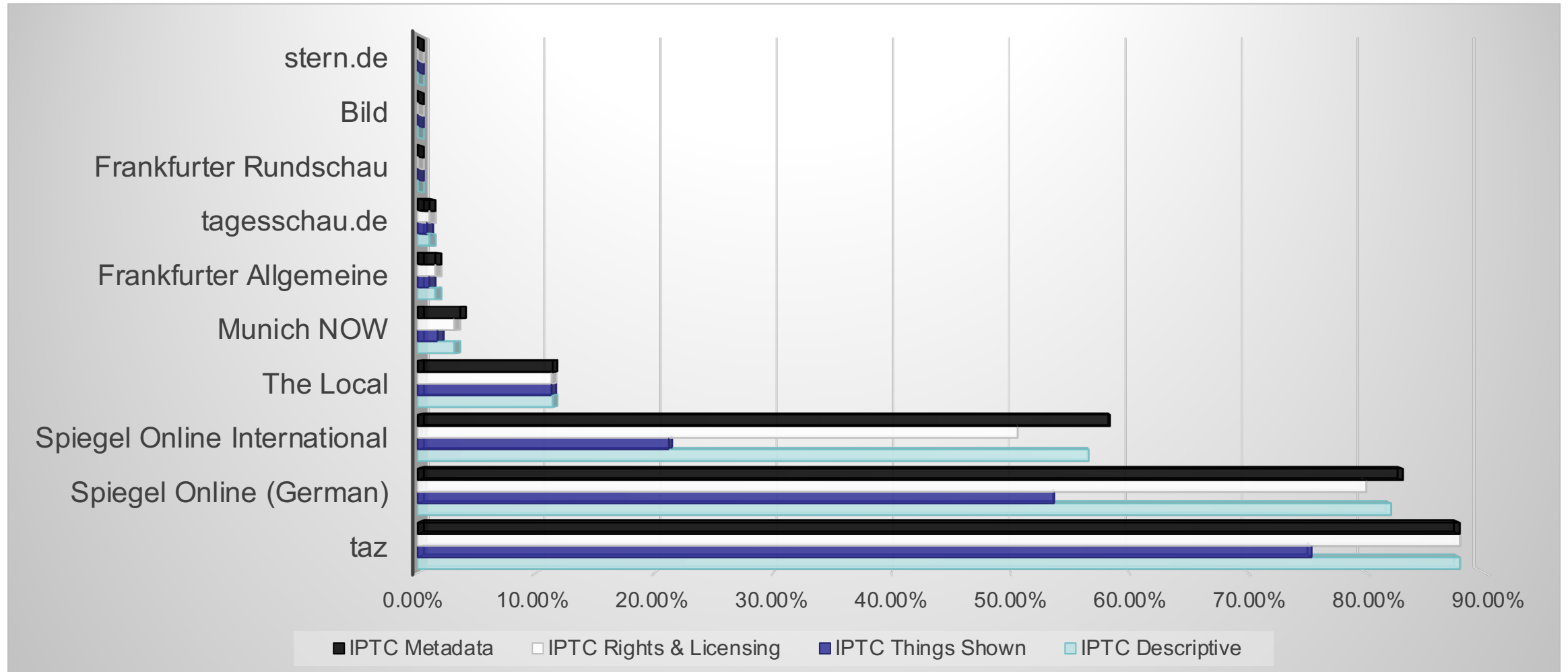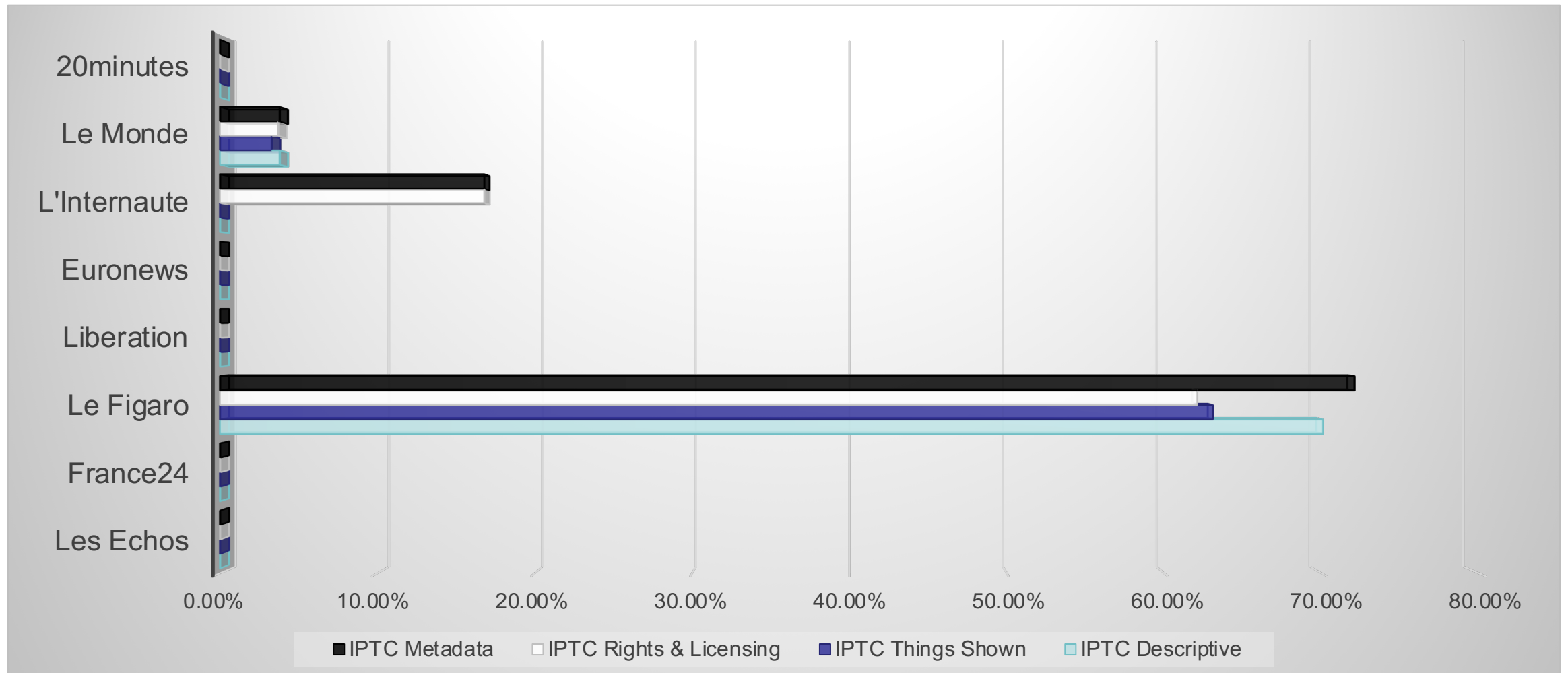| Rank	Country name	Number of feeds	IPTC Descriptive	IPTC Things Shown	IPTC Rights & Licensing	IPTC Metadata
16	Spain	6	0.50%	0.00%	0.50%	0.50%
17	Poland	3	0.80%	0.40%	0.40%	0.80%
18	Portugal	5	0.50%	0.40%	0.40%	0.50%
19	Estonia	6	0.30%	0.00%	0.30%	0.30%
20	Lithuania	4	0.20%	0.00%	0.20%	0.20%
21	United Kingdom	21	0.10%	0.00%	0.20%	0.20%
22	India	5	0.10%	0.10%	0.10%	0.10%
23	Romania	3	0.10%	0.10%	0.10%	0.10%
24	Sweden	5	0.10%	0.10%	0.10%	0.10%
25	Belgium	3	0.40%	0.00%	0.00%	0.40%
26	Bulgaria	3	0.00%	0.00%	0.00%	0.00%
27	Croatia	4	0.00%	0.00%	0.00%	0.00%
28	Czech Republic	6	0.00%	0.00%	0.00%	0.00%
29	Finland	3	0.00%	0.00%	0.00%	0.00%
30	Ireland	4	0.00%	0.00%	0.00%	0.00%
31	Malta	3	0.00%	0.00%	0.00%	0.00%
32	Netherlands	4	0.00%	0.00%	0.00%	0.00%
33	Norway	6	0.00%	0.00%	0.00%	0.00%
34	Slovakia	3	0.00%	0.00%	0.00%	0.00%
35	Switzerland	2	0.00%	0.00%	0.00%	0.00%

- 20 countries (of 35 studied) have little or no embedded metadata
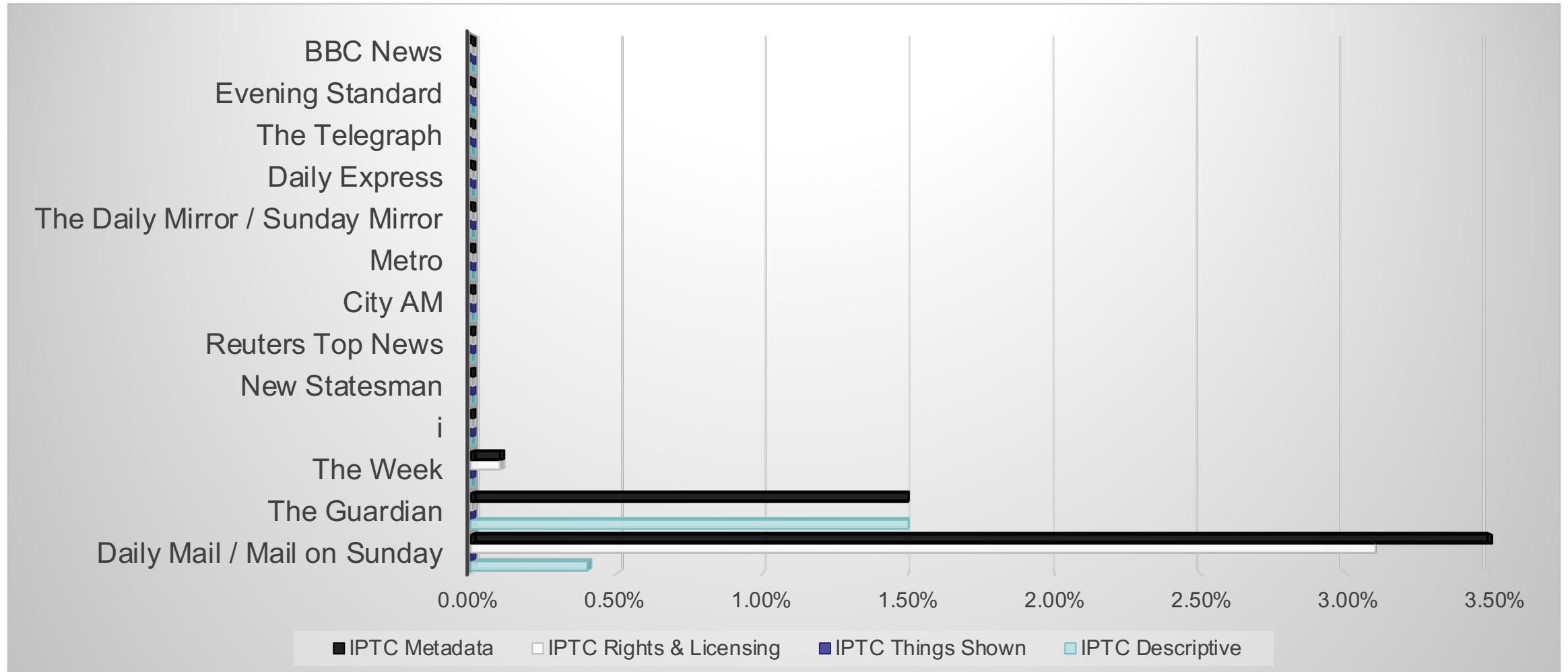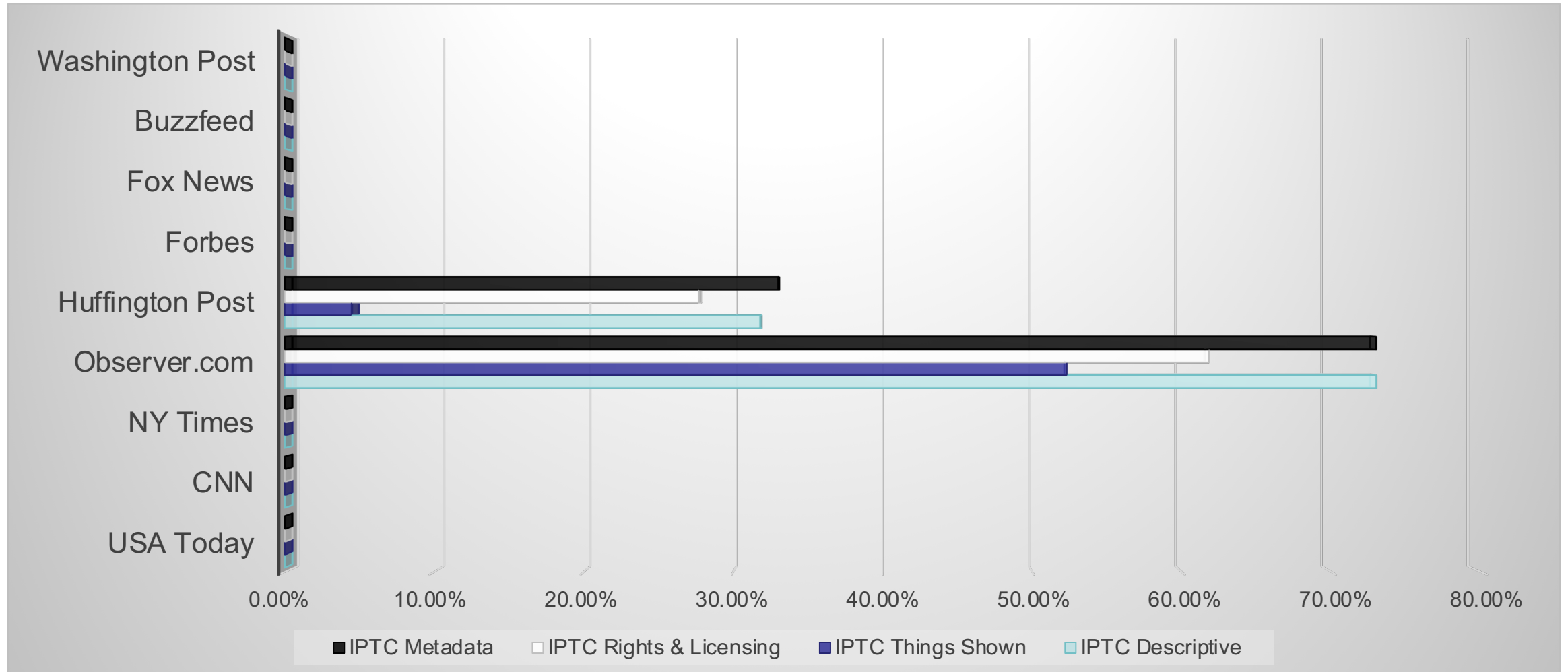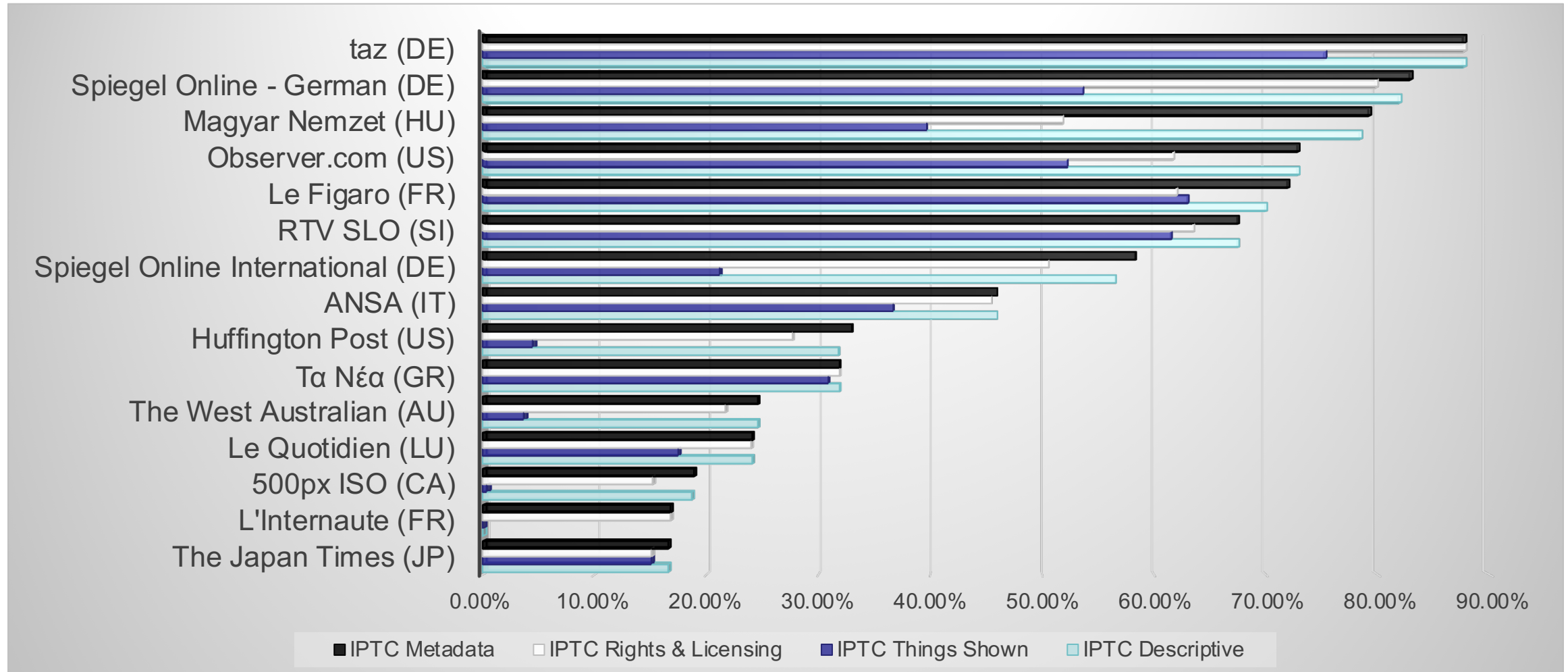- We have our work cut out for us!

Feeds in Germany

Legend: IPTC Metadata ■ IPTC Rights & Licensing □ IPTC Things Shown ■ IPTC Descriptive ■

Categories (top to bottom): stern.de, Bild, Frankfurter Rundschau, tagesschau.de, Frankfurter Allgemeine, Munich NOW, The Local, Spiegel Online International, Spiegel Online (German), taz

X-axis: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%, 90.00%

Feeds in France

Legend: IPTC Metadata, IPTC Rights & Licensing, IPTC Things Shown, IPTC Descriptive

Categories (top to bottom): 20minutes, Le Monde, L'Internaute, Euronews, Liberation, Le Figaro, France24, Les Echos

X-axis: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%

Feeds in United Kingdom

Chart: horizontal bar chart with vertical axis (top to bottom): BBC News, Evening Standard, The Telegraph, Daily Express, The Daily Mirror / Sunday Mirror, Metro, City AM, Reuters Top News, New Statesman, i, The Week, The Guardian, Daily Mail / Mail on Sunday. Horizontal axis: 0.00%, 0.50%, 1.00%, 1.50%, 2.00%, 2.50%, 3.00%, 3.50%.

Legend: ■ IPTC Metadata □ IPTC Rights & Licensing ■ IPTC Things Shown ■ IPTC Descriptive

Feeds in United States

Legend: IPTC Metadata, IPTC Rights & Licensing, IPTC Things Shown, IPTC Descriptive

Categories (top to bottom): Washington Post, Buzzfeed, Fox News, Forbes, Huffington Post, Observer.com, NY Times, CNN, USA Today

X-axis: 0.00%, 10.00%, 20.00%, 30.00%, 40.00%, 50.00%, 60.00%, 70.00%, 80.00%

And now for the final results...

Overall top feeds globally

Conclusions

- A lot of education is needed

- Hopefully publishing some of this information will encourage publishers to treat embedded metadata more kindly

- Better tools are still needed

- Awareness will drive better support in DAMs, CMSs, CDNs

- More documents like the "Quick Guide to IPTC Photo Metadata and Google Images" could help

Next steps

- Publish initial results on iptc.org blog
- Make the browser available to IPTC member companies
- Add some features:
  - Month-by-month comparisons over time
  - Get better at only extracting editorial images (we're currently grabbing some incidental images e.g. from Outbrain panels)
  - Compare usage of particular fields across publishers and countries
- Add more feeds, more countries
- All suggestions greatly received!

Please get in touch!
Twitter: @brendanquinn / @IPTC
Email: mdirector@iptc.org