IIPC 2019

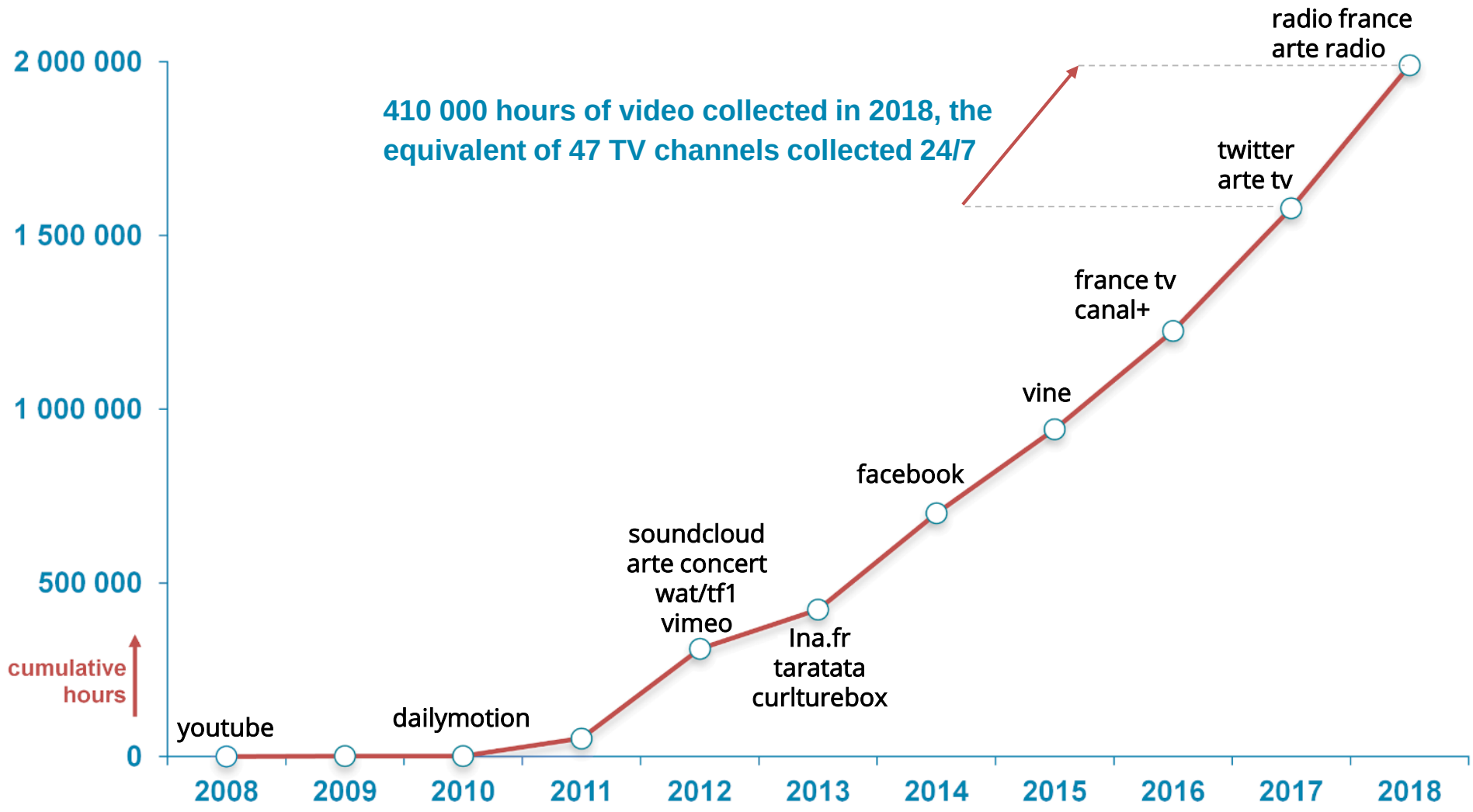
From videos to channels:
archiving video content on the web

-

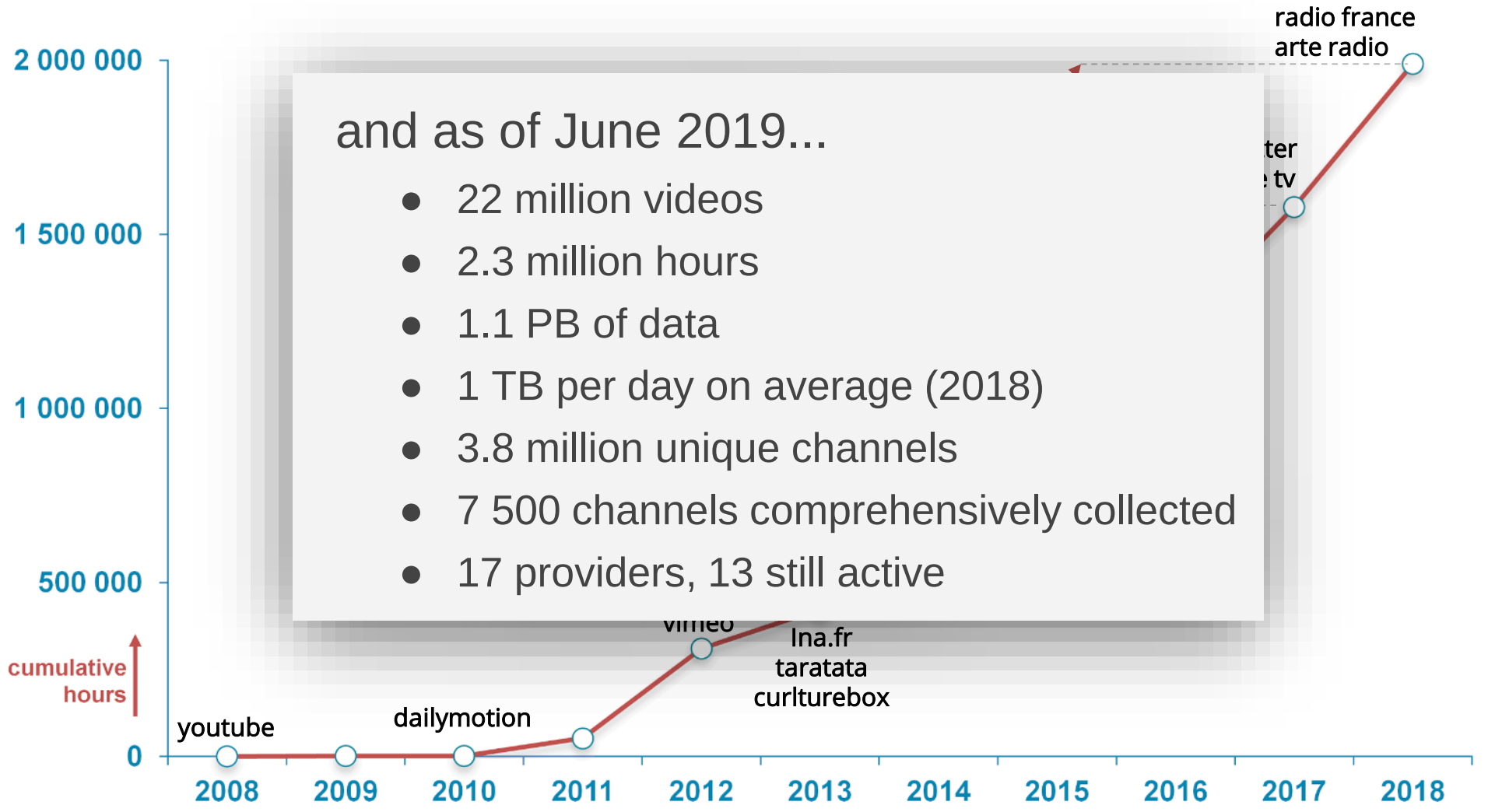
Thomas Drugeon, INA

ina

Video capture evolution @INA



Video capture evolution @INA



and as of June 2019...

- 22 million videos
- 2.3 million hours
- 1.1 PB of data
- 1 TB per day on average (2018)
- 3.8 million unique channels
- 7 500 channels comprehensively collected
- 17 providers, 13 still active

Video URN

- Uniquely identifying videos and channels for sourcing, crawling and access purposes
- Part of our global dlweb URN scheme
- Videos
  - `web:av:youtube:jNQXAC9IVRw`
  - `web:av:twitter:560070131976392705`
- Channels / Users / Accounts / Authors
  - `web:user:facebook:france2`
  - `web:user:youtube:UCPlxmWbGbwVM2baH8q7MK6w`

Video sourcing

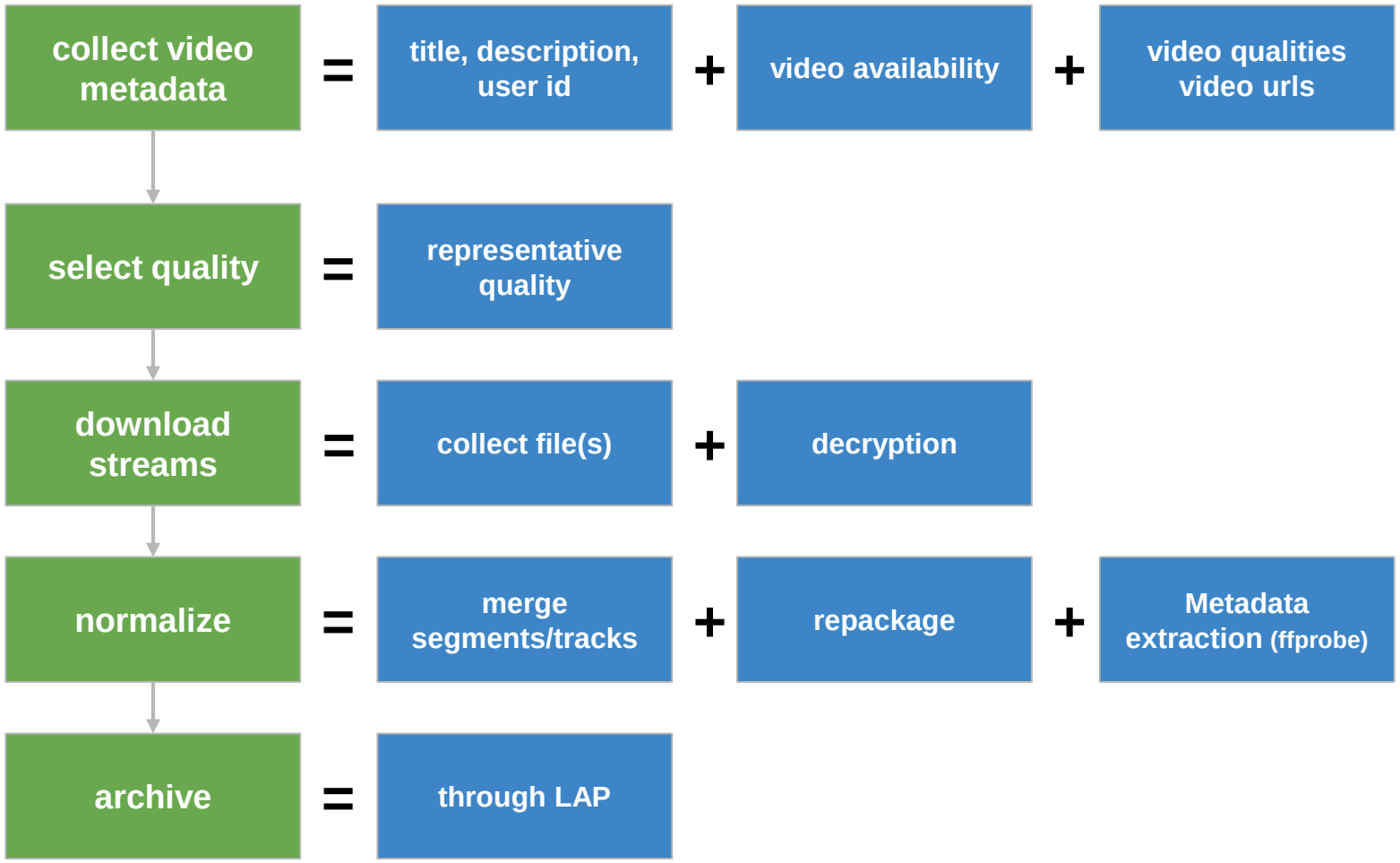
- Direct listing of publication for a given channel
  - From 7500 manually sourced channels
    (TV/radio related or native)
  - Specific to each provider: APIs, feeds, scraping, ...
- Continuous archive recrawl
  - DAFF-tool detecting video ids in feeds and HTML pages
  - Also searching for video embedded in tweet objects

→ URN list sent to video capture system

Video capture

- In-house dedicated crawling system with specific code for each platform (DevOps)
- Each unique video URN is crawled once
- Around 1TB per day on avg (2018)
- Data selection (format, resolution, ...)
- Metadata normalization
- Archive with the LAP (Live Archiving Proxy)
- DAFF storage (1.1PB)

UGC Monitor: download process

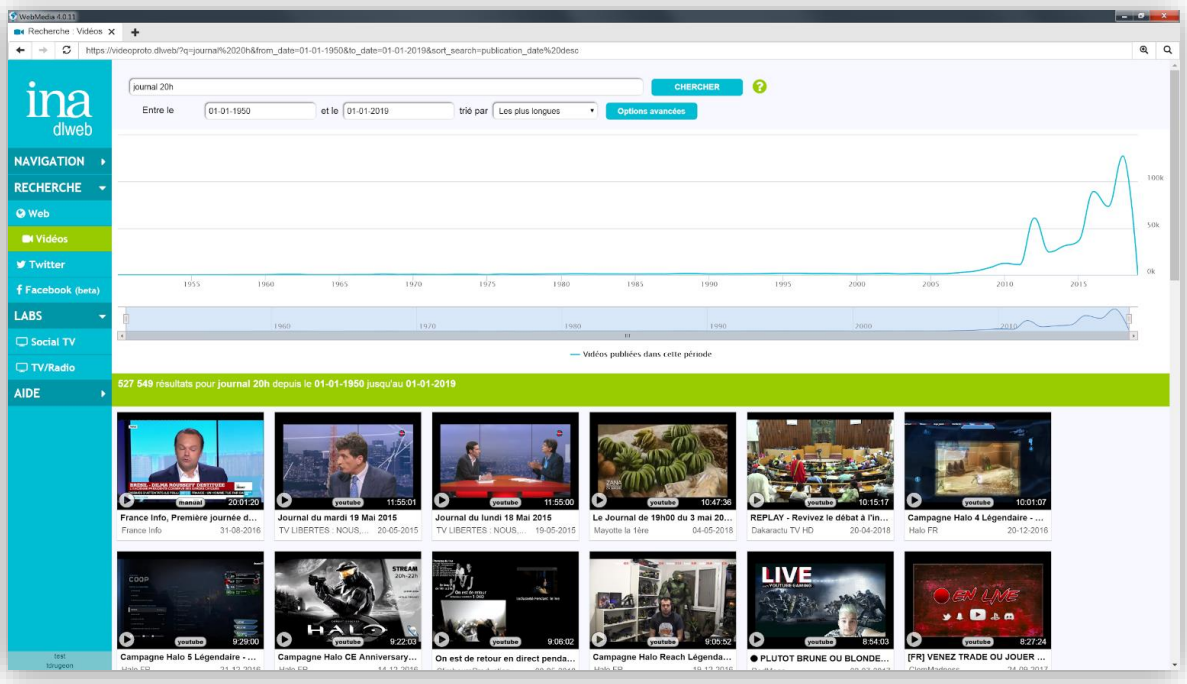


Access

- URN identification
- Dedicated streaming engine
  - Same webservice and player for all INA medias (TV/radio have different URN namespaces)
  - On-the-fly conversion (mp4, flv, webm, ogg, ...)
  - On-the-fly extraction (image, preview, waveform, ...)
- Integration
  - Web archive browsing (video id detection)
  - Twitter & Facebook search
  - Dedicated video search engine (metadata)
- `<dlweb-preview-video urn="...">` web component

Access Demo

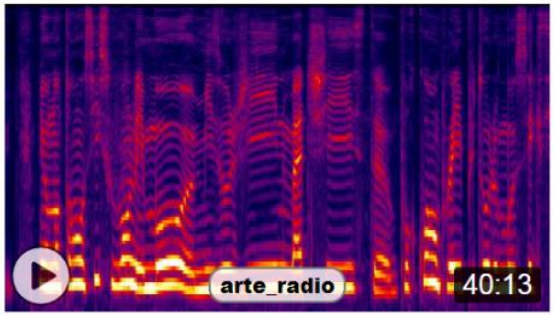
https://youtu.be/qlOY8dWSiMA



Video preview web component

web:av:youtube:i7ueOkO2EFY

web:av:arte_radio:61659100



web:av:facebook:831043176969172

flux:tv:fr2:20190429T202000:2700

Issues and future considerations

- Obsolete video codec (eg vivo) calling for batch conversions
- Best-effort and platforms goodwill
- DRM even for free contents: rationalisation
- youtube-dl