Flexible data scraping, multi-language indexing, entity extraction and taxonomies: Tadam, a Swiss tool to deal with huge amounts of unstructured data

Titus Plattner
Investigative reporter, Tamedia
Bern, Switzerland
titus.plattner@tamedia.ch

Didier Orel
IT-Publishing, Tamedia
Lausanne, Switzerland
didier.orel@tamedia.ch

Olivier Steiner
Business development, Tamedia
Zurich-Lausanne, Switzerland
olivier.steiner@tamedia.ch

ABSTRACT

In this paper, we describe an innovative data storing and sharing system, developed since 2015 in the Swiss news organization Tamedia. The system automates and centralizes information gathering tasks and makes collaborative work easier.

Category: Data Management; Data Preservation; Data Archiving; Computational Journalism; Entity relationship modeling ; Web crawling; Natural language processing; Information extraction; Machine translation; Lexical semantics; Semantic networks
Keywords: Media, Collaborative work, Leaks, Data mining, Investigative journalism, Public-interest reporting, Government accountability, Innovation project, Taxonomies

1. INTRODUCTION

News organizations face an increasingly complex world, which produces literally a permanent data flood. In the meantime, most of them struggle with limited and often declining resources for in-depth journalism [1]. Being able to manage huge amounts of data and a collaborative approach have been the keys to the success of the International Consortium of Investigative Journalists (ICIJ) investigations during the last three and a half years, from Offshore Leaks to Panama Papers [2]. One of the authors of this paper was part of these investigations in Switzerland. He proposed to implement ICIJ's scheme in Tamedia, his news organization, not only in situations of big leaks, but also in the company's daily life. This is the goal of an internal project named Tadam, financed by the company's innovation fund. The total budget is $1.5 million for 2016-2018.

One of the challenges is that the users are heterogeneous. Tamedia is a Swiss news organization employing more than 3,500 people. One third of them are journalists, working in about 40 media outlets. Some of those are very local, others, like the *Tages-Anzeiger*, are leading national newspapers, competing on a European level. But at every branch of the company, the digitalization has completely changed the amount of data that are accessible to journalists over the past 20 years. When before, Tamedia's reporters waited for 5 or 10 pages of documents arriving per fax to write an article, they now often have access to hundreds of pages of reports, raw documents, court decisions, etc. But in the meantime, how documents are managed hasn't really changed. The consequence is that in most newsrooms, reporters waste data that they have but don't use in their own reporting.

The Tadam project has three goals: motivate journalists to store their information for future reports and to share it if possible, in small teams or on the company level; make that data accessible for their colleagues, even after the single reporter who acquired it is not present or has left the company; and automate and centralize as much as possible information gathering tasks. The idea behind the system is not to build a very organized archive, but rather to base it on a powerful search engine.

Technical features of the system that help the journalists to exploit the maximal potential of their growing data resources include geographical search, smart alerts, and name extraction. This allows them not only to search the data, but provides the ability to explore it.

In an article about how computers can empower public-interest journalism published in 2011 [3], Sarah Cohen, James T. Hamilton and Fred Turner identified five areas of opportunity:

- Combining information from varied digital sources (e.g., searching websites for news of indictments);
- Information extraction (e.g., entity extraction);
- Document exploration and redundancy (e.g., finding what's new, and mining accumulated documents);
- Audio and video indexing;
- Extracting data from forms and reports (e.g., translating tables of data in pdfs into excel);

Based on their intuition and on Tamedia's needs, the authors have started to implement a new single tool that tackles the three first challenges listed.

2. WHAT IS TADAM?

Tadam, which is the abbreviation of *Tamedia Data Mining Project,* is a web-based private platform that securely collects and stores massive amounts of unstructured data and makes it easily searchable, regardless of the format or language of the original source. The system is built on a software core of an external company, and adapted for a Swiss news organization. Expert System, the Italian software supplier, had no experience working with news organizations. The software is mainly used by oil or insurance companies and State agencies, essentially for social network monitoring. The Tadam platform is still in a beta version.

During the last twelve months Expert System and Tamedia have been customizing it.

To optimize the utility-cost ratio, it is not only designed for large datasets coming in at once (leak situations), but also for the daily specialized reporting (journalists continuously gathering smaller amounts of curated data, creating their own digital archive), and even for live reporting (data extracted automatically directly online).

3. FLEXIBLE DATA ACQUISITION

Basically, journalists receive documents or official information from many different channels: they get press releases via e-mails, they gather publications from the internet, they subscribe to RSS feeds, they follow Twitter or Facebook accounts, they still receive a lot of information on paper, sometimes, they are blessed with a bunch of documents on a USB stick or a hard drive. Tadam is designed to make all these different signals searchable and cross-referenceable from a single entry point.

This is why the acquisition workflow is so important. Let's distinguish *static sources*, that the users (journalists) put in the system manually; and *dynamic sources*, which are targets, that the system has to monitor on its own, and grab when something new becomes available.

3.1 Dynamic sources

Dealing with dynamic sources is quite challenging from a technical point of view. There are different systems available on the market, but we believe, Tadam offers the broadest range of possible source types. This flexibility is really important, because you can't ask journalists only to use sources that are publishing information in a certain format.

The information they need is often published on dirty coded websites, that avoid any clean data extraction, or sent by e-mails containing attachments in a format you first have to OCR.

Tadam is able to verify every ten minutes if a website containing links to pdf documents had an update and to download only these pdfs. The setup of the system's web crawling is quite intuitive and manageable in most cases directly by journalists after a short introduction. Supported source types are: Webpages, RSS, Twitter, News (Google, Yahoo, Bing), Search (Google, Yahoo, Bing), e-mail.

During the test phase, users from the *Berner Zeitung*, one of Tamedia's regional newspapers, asked for the capability to push SMS into Tadam. The reason was that in the Swiss canton of Bern, high river levels and flooding alerts are only sent via short text messages. With a $30 Android phone, we quickly built a connector, relaying those SMS via e-mail into the system. This anecdote illustrates the importance of multiple scraping methods. In the future, we will probably add other source types.

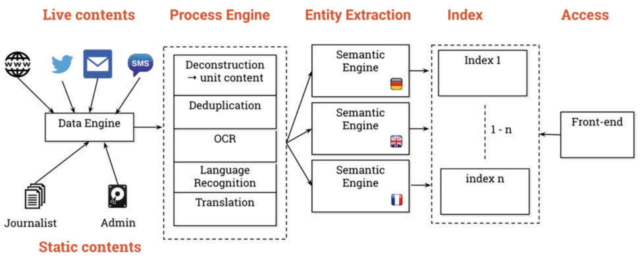


Figure 1. Tadam's acquisition workflow (simplified).

3.2 Static sources

For static sources, the number of file formats the system has to deal with is a bit higher. Tadam is not an IT forensic analysis tool like Nuix, but it has dealt with more than 30 of the most common formats (.pdf, .doc, .ppt, rtf, xls, .zip, etc). Users can upload their files into the system and the files are automatically processed.

In a pilot phase, we pushed some folders from the Panama Papers into the system. There were thousands of e-mails, often containing attachments that had to be processed automatically. The system was able to extract all those sub files, up to a theoretically unlimited level, while keeping the link to the parent file (Deconstruction phase).

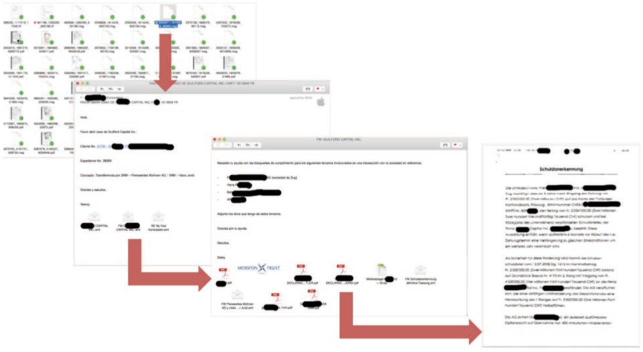


Figure 2. : E-mails from the Panama Papers, containing other e-mails as attachment, that contain .xls and .pdf documents, with some to OCR.

In case of reasonable amounts of data, the system is able to ingest those files smoothly. But if hundreds of Gigabytes (or Terabytes) quickly need to be processed (big leak situations), a preprocessing is the best solution. The unnecessary steps of the standard processing are spared, in order to maximize the speed.

Note that the duplicates are not acquired twice. The system recognizes them by using a simple MD5 hash of the different files. In a dataset consisting of a copy of a leaked mail server we analyzed for an ongoing project, we had about 40% duplicates.

4. OCR AND TRANSLATION

The next steps of the processing are OCR and translation, managed for more than 60 source languages (from Indo-European languages like Spanish or Russian, to Chinese, Arabic or Hindi) into three working languages (German, French and English).

When journalists have paper documents to OCR, they can send them into the system directly from the photocopier or their smartphone. The system uses the ABBYY OCR-engine, which has in our experience the best output quality.

Once the text is available, the language of the document is automatically recognized and translated in the different working languages. For the most common languages, we use the Systran translation-engine installed on our own servers. For about 5% of the documents that are in exotic languages, we use Google translate. This translation phase improves the usability of the system, especially for Tamedia's users, who work in a multilingual context, dominated by German and French. Firstly, the translation allows quick evaluation of the content of a document, even if it is in a non-familiar language. Secondly, it facilitates finding documents with a single search, regardless of their language.

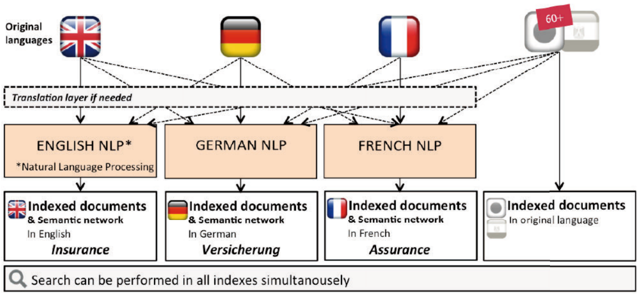
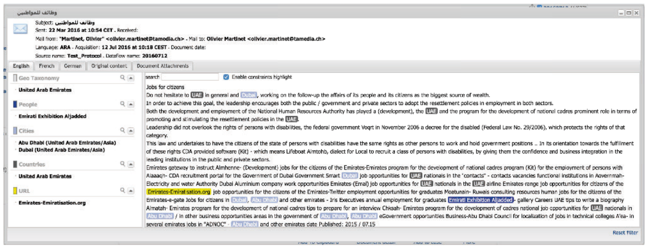

Original languages
60+
Translation layer if needed
ENGLISH NLP*
*Natural Language Processing
GERMAN NLP
FRENCH NLP
Indexed documents & Semantic network In English Insurance
Indexed documents & Semantic network In German Versicherung
Indexed documents & Semantic network In French Assurance
Indexed documents In original language
Search can be performed in all indexes simultaneously

Figure 3. The dealing with different languages (simplified).

With a single search "insurance" and "Switzerland" it is for example possible to find documents containing the equivalents in German ("Versicherung" and "Schweiz") or French ("assurance" and "Suisse"). With other search tools, it is often necessary to repeat the same search in two or three languages. The extracted text of the document can be displayed in different tabs for each language.



وظائف للمواطنين

Subject: وظائف للمواطنين
Sent: 22 Mar 2016 at 10:54 CET . Received:
Mail from: "Martinet, Olivier" <olivier.martinet@tamedia.ch> - Mail to: Olivier Martinet <olivier.martinet@tamedia.ch>
Language: ARA - Acquisition: 12 Jul 2016 at 10:18 CEST - Document date:
Source name: Test_Protocol . Dataflow name: 20160712

Geo Taxonomy
- United Arab Emirates

People
- Emirati Exhibition Aljadded

Cities
- Abu Dhabi (United Arab Emirates/Asia)
- Dubai (United Arab Emirates/Asia)

Countries
- United Arab Emirates

URL
- Emirates-Emiratisation.org

search ☑ Enable constraints highlight

Jobs for citizens
Do not hesitate to UAE in general and Dubai, working on the follow-up the affairs of its people and its citizens as the biggest source of wealth.
In order to achieve this goal, the leadership encourages both the public / government and private sectors to adopt the resettlement policies in employment in both sectors.
Both the development and employment of the National Human Resources Authority has played a (development), the UAE and the program for the development of national cadres prominent role in terms of promoting and stimulating the resettlement policies in the UAE.
Leadership did not overlook the rights of persons with disabilities, the federal government Voqrt in November 2006 a decree for the disabled (Federal Law No. 29/2006), which protects the rights of that category.
This law and undertakes to have the citizens of the state of persons with disabilities have the same rights as other persons to work and hold government positions .. In its orientation towards the fulfillment of these rights CDA provided software (Kit) - which means Lifeboat Almtohb, dialect for Local to recruit a class of persons with disabilities, by giving them the confidence and business integration in the leading institutions in the public and private sectors.
Emirates gateway to instruct Almhenne- (Development) jobs for the citizens of the Emirates-Emirates program for the development of national cadres program (Kit) for the employment of persons with Alaaaqh- CDA recruitment portal for the Government of Dubai Government Smart Dubai job opportunities for UAE nationals in the 'contacts' - contacts vacancies functional institutions in Aovernmah- Electricity and water Authority Dubai Aluminium company work opportunities Emirates (Emal) job opportunities for UAE nationals in the UAE airline Emirates-range job opportunities for citizens of the Emirates-Emiratisation.org job opportunities for the citizens of the Emirates-Twitter employment opportunities for graduates Moatunain- Ruwais consulting resources human jobs for the citizens of the Emirates-e-gate Jobs for citizens in Dubai, Abu Dhabi and other emirates - Iris Executives annual employment for graduates Emirati Exhibition Aljadded- gallery Careers UAE tips to write a biography Almatah- Emirates program for the development of national cadres tips to prepare for an interview Chksah- Emirates program for the development of cadres national job opportunities for UAE nationals in Abu Dhabi / in other business opportunities areas in the government of Abu Dhabi, Abu Dhabi eGovernment opportunities Business-Abu Dhabi Council for localization of jobs in technical colleges A'ea- in several emirates jobs in "ADNOC" - Abu Dhabi and other emirates date Published: 2015 / 07.15

Figure 4. The translated version of an Arabic document, with the extracted entities highlighted.

5. ENTITY EXTRACTION

The system is able to extract entities like names, company names, organizations, places, cities, countries, and also phone numbers, e-mail addresses, etc. Each natural language processing engine (NLP) from the three system languages works separately. The entity extraction is a heuristic process and is not 100% accurate, especially after the alteration of the automatic translation. In order to improve the ratio of recognized entities, we use the following method: if two NLP's detect the same entity, this entity is injected in the third language where it wasn't detected.

Those extracted entities show in a very efficient manner what's inside a document set, instead of searching what could be inside. And it provides to the journalist the ability to explore the data. Especially in a big leak situation, this approach saves a lot of time, and the risk of missing something important decreases significantly.

Geographical entities receive coordinates and can therefore be displayed on a map, in order to search for documents about a certain area.

At the end, the entity extraction allows the journalist to extract structured data from an unstructured dataset. In the pilot phase, we used for example the system to analyze a set of more than 800 .pdfs from the Swiss federal Gazette about administrative assistance on tax matters. Within 20 minutes, the system produced a statistic of how many publications were related to which country, and all the people named in these publications.

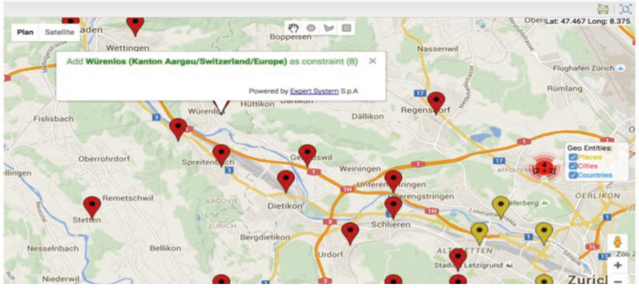


Figure 5. Exploring the documents on a map.

To be absolutely sure, we opened all the 806 .pdf files manually and counted every entry. It took two people three hours. There was only one discrepancy: the system identified the former French president's name Georges Pompidou, extracted from a street address "rue Georges Pompidou." Of course, the deceased former French president was not the subject of a request on tax matters.

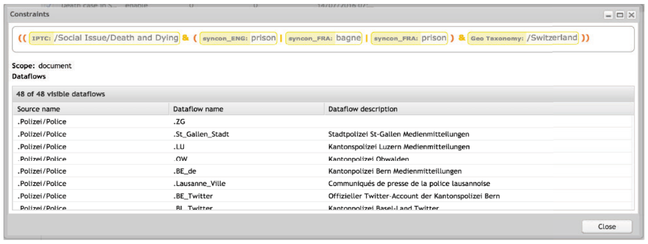
6. TAXONOMIES

The system also works with taxonomies. A taxonomy search with "vehicle" defined as a "conveyance that transports people" will find the synonyms and also documents containing all subcategories like "car", "train", "motorcycle". In the meantime, it won't find documents containing the synonyms of "vehicle" as an "agent" like "factor", "medium", etc.

Geo taxonomies allow one to find a document containing "California" and "fraud" with a search "United States AND fraud", because the system knows that "California" is a subcategory of "United States". For Switzerland, we added a geographical layer up to the third order administrative division, which is the municipality level. And we consider, in a further phase, to add the much more detailed SwissNAMES3D (300,000 entries), which contains, not only points, but also lines (rivers, roads, etc.) and polygons (surface of a district, of a lake, etc.). Tadam also automatically sets IPTC taxonomy codes to each document.

7. COMPLEX QUERIES AND ALERTS

It is very important to be able to do complex queries in the search engine. Tadam is of course able to search with Boolean operators and entity descriptors, and combining them with commas. This allows for example to search efficiently every police press release or court decision about a death case in a Swiss prison. In Switzerland, the judiciary system is very decentralized: 26 cantons with 26 separate police agencies, prosecutorial offices and carceral administrations. A central databank about incidents in prisons doesn't exist.



Constraints

((IPTC: /Social Issue/Death and Dying & (syncon_ENG: prison | syncon_FRA: bagne | syncon_FRA: prison) & Geo Taxonomy: /Switzerland))

Scope: document
Dataflows

48 of 48 visible dataflows

Source name	Dataflow name	Dataflow description
.Polizei/Police	.ZG	
.Polizei/Police	.St_Gallen_Stadt	Stadtpolizei St-Gallen Medienmitteilungen
.Polizei/Police	.LU	Kantonspolizei Luzern Medienmitteilungen
.Polizei/Police	.OW	Kantonpolizei Obwalden
.Polizei/Police	.BE_de	Kantonpolizei Bern Medienmitteilungen
.Polizei/Police	.Lausanne_Ville	Communiqués de presse de la police lausannoise
.Polizei/Police	.BE_Twitter	Offizieller Twitter-Account der Kantonspolizei Bern
.Polizei/Police	.BL_Twitter	Kantonpolizei Basel-Land Twitter

Figure 6. A complex search for the press releases about death cases in Swiss prisons.

8. SECURITY AND ARCHITECTURE

For better usability and greater scalability, the system is web based. The data and the processing happen on premises, on the company's own servers. The client server communications are encrypted and a two-factor authentication is required to access the system.

The goal is to give access to every journalist inside the company. The architecture is built to allow a rapid increase of users and a continuous increase of stored data.

Law enforcement authorities use similar systems, but once a case is archived, the data on the analysis platform has to be retrieved. And in general, it is not permitted to cross reference different cases. In contrast, the information packages stored inside Tadam won't be disabled. The more data there is active on the servers, the higher the chance there will be to find critical information for the reporting.

In order to encourage the journalists to use the system and to put their own high-value information inside, it will be necessary to be flexible with the access rights to account for every situation. Users can keep their data private, share it with individuals or smaller groups, or open it to every Tadam user, inside the company.

When the journalist leaves, he has the possibility to take with him the raw data collected by him. If the data was private, no copy will be kept on the platform. If the data was shared, the other users will continue to have access to it, in order to keep this potentially highly valuable information inside the company. This is a long-term investment, especially in journalistic fields where the online data accessibility is lower, like local journalism.

Very deliberate management of access rights and high IT security are the only ways to guarantee ethically responsible behavior of the news organization regarding source protection and respect of privacy.

9. EXPECTATIONS

Tadam is an internal innovation project and the system is still in a beta phase. Currently, the platform has 40 users. Primarily, it is a project that should allow a quality improvement in journalistic output. Possible productivity gains are not the driving force of the project.

The core team of the project imagined different ways to use the system but it is expected that the users will invent other applications. Here are some examples:

9.1 Leak processing

During the last five years, the number and the volume of documents leaked to journalists has increased significantly. Since the Afghanistan War Logs of Wikileaks in July 2010, there was an acceleration of very big data leaks: US diplomatic cables (Nov. 2010), Gitmo Files (April 2011), Syria Files (Oct. 2012), Offshore Leaks (April 2013), Snowden Files (June 2013), Chinaleaks (Jan. 2014), Yanukovychleaks (Feb. 2014), LuxLeaks (Nov. 2014), Swissleaks (Feb. 2015), Sony Archives (April 2015), Kazakhstan Files (May 2015), Saudi Cables (June 2015), Hacking Team (July 2015), Drone Papers (Oct. 2015), etc. This phenomenon, mainly visible on the international level, has also broken out on a national and regional level in Switzerland. Among other files, Tamedia's investigative unit for example received more than 23,000 pages of justice documents about the Falciani case, and an internal mail server of a Geneva based oil-trading company.

Every large news organization that wants to play a leading role in the investigative field has to be able to process such "Terabyte big" leaks of unstructured data. To our knowledge, only a very few news organizations are able to do this. Non-profit organizations like ICIJ or OCCRP, which are able to mix the journalistic and the coder cultures, are pioneers in this field.

In terms of functionalities, Tadam is much more advanced than Blacklight, ICIJ's platform that was built with open source bricks by its very small team of developers. But in terms of efficiency, reaction speed and resistance to a higher amount of users, Tamedia's system quality has yet to be proven in the field.

9.2 Traditional reporting

The system is also designed for daily usage in traditional reporting. The amount of information gathered by journalists is extensive, yet they are not able to access it efficiently with traditional archiving systems on paper or on computer hard drives. Remote access on a server like Tadam corresponds to their new mobility habits. Information will be easier to find, even years later.

Often different members of the same news organization may possess individual information puzzle pieces but their colleagues are unaware they have them. Collaborative approaches in journalism have proven their efficiency. Exactly like cross-border journalism, Tamedia wants to promote "cross-regional" and section transversal collaborations to improve the quality of reporting. The ability to share and to work on same document sets (creating collections, adding notes, rating, etc.) introduced with Tadam will help to reach this goal.

The particular multilingual situation of Switzerland (German (64%), French (23%), Italian (8%)), with three media landscapes that pulse in different rhythms, is a very interesting place to try this new approach. The systematic translation of the acquired documents should help to break the language barriers.

Today, in Switzerland, press officials from different administration units are working in coordination. If a journalist asks for something from the Swiss Air Force, the communication division of the Defense Department is immediately informed, through a new information management system. But journalists rarely work together. It can happen that two reporters get contradictory responses from those same press officials. We think that sharing on Tadam all the official e-mails in certain fields could be an efficient counter-measure to the increasing influence of communication specialists.

Sharing systematically on the news organization level the documents obtained after a long FOIA procedure can also reinforce public-interest reporting and government accountability.

The scraping functionality of the system allows the monitoring of hundreds of websites in order to alert reporters when something relevant is published. On websites that publish something new only once every several weeks, reporters wouldn't have the patience or the discipline to check "manually" every day, in order to be the first who notices the change.

This usage can be a good answer to the following problem: public pressure has forced States to be more transparent and publish more documents by themselves. But often, this is a façade of transparency, and the documents are discretely published, without any notification to journalists, on subsections of websites that are very difficult to find. The Swiss Federal Statistical Office often uses this trick.

A report published in April 2016 by the State chemists office of the canton of Valais contained explosive information: every fourth analyzed cheese contained a high level of *Escherichia coli* – the fecal bacteria. Nobody noticed until an intern colleague finally found the report in July. The Cantonal chemists did not hold a press conference or generate a media release. Now, with Tadam, a reporter can monitor this subsection of the website and be alerted automatically as soon as the new publication is placed online.

As long as the websites do not contain a blocker for search engines, it is possible to put in place similar monitoring by using Google Alerts or other tools. But you can't be sure to get the alert 10 minutes or one hour after the change happened. Websites with low traffic are indexed by Google only every several weeks or months.

9.3 Live reporting

Live reporting is faster thanks to real time information acquisition from the Internet. Today, Tamedia's newsrooms have no real structure for primary information gathering. Some have a central e-mail account, used by the front desk, others, just rely on different journalists. Not every newsroom subscribes to every information issuer.

For example, if an event with several death cases happens in the city of Winterthur, the journalists at the front desk at the *Tribune de Genève* will only get this information indirectly, via the Swiss news agency or another news outlet. With Tadam, this process is centralized and accelerated. In a test done in the pilot phase of the project, the press releases of the 26 Cantonal police agencies (scraping frequency set up every 10 minutes) were acquired on average 45 minutes before the fastest news report from a news website.

For web scouters at the front desks, this tool allows a scale change. Instead of monitoring 30 or 50 websites, mainly of competitors, it is possible to follow one central feed bringing together 500 primary and secondary sources.

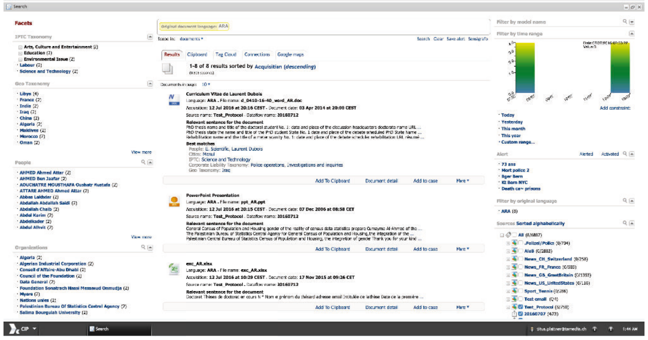
In situation of a breaking news event, it is possible to set up new web sources quickly or Twitter hash tags to scrap, etc. Just after the Terror attack in Nice on July 14, 2016, it was possible to crawl local news websites, relevant twitter accounts and the press release section of the local police website.

The live reporting is also improved because reporters have access to the history of the crawled sources. If set on collections of primary sources, the granularity of the information is much higher than media archive databases, which often comprise the main historical source for journalists.

This continuously growing historical information adds context to live reporting, for example, by explaining that a deadly electric bicycle accident is the twenty-second this year, twice as much as in 2015. Even if it was deleted on the internet, the information remains stored in Tadam. The system is also able to grab two different versions of the same online publication. This allows one to see what words have been changed between two different versions of an official communication, which is often rich in information.

10. CHALLENGES

Working with a system like Tadam requires not only a short training, but also close individual coaching, in order to guide the change of mindset it requires. People at the front desk have to acquire new reflexes. And the attitude change, from lone wolf habits to a transparent collaboration, is a huge step for many reporters. In order to facilitate this, patient persuasion work is necessary. We know that this will be the most difficult part of the project. Middle management has to be involved to conduct this change.



1-8 of 8 results sorted by Acquisition (descending)

Curriculum Vitae de Laurent Dubois
Language: **ARA** - File name: **d_0410-16-40_word_AR.doc**
Acquisition: **12 Jul 2016 at 20:16 CEST** - Document date: **03 Apr 2014 at 20:00 CEST**
Source name: **Test_Protocol** - Dataflow name: **20160712**
Relevant sentence for the document
PhD thesis name and title of the doctoral student No. 1: date and place of the discussion headquarters doctorate name URL ...
PhD thesis state the name and title of the PhD student State No. 1 date and place of the debate scheduled PhD State Name ...
Rehabilitation name and the title of a meter scanty No. 1: date and place of the debate scheduler rehabilitation URL résumé ...
Best matches
People: E. Scientific, Laurent Dubois
Cities: Mosul
IPTC: Science and Technology
Corporate Liability Taxonomy: Police operations, investigations and inquiries
Geo Taxonomy: Iraq

PowerPoint Presentation
Language: **ARA** - File name: **ppt_AR.ppt**
Acquisition: **12 Jul 2016 at 20:15 CEST** - Document date: **07 Dec 2006 at 08:58 CET**
Source name: **Test_Protocol** - Dataflow name: **20160712**
Relevant sentence for the document
General Census of Population and Housing gender of the reality of census data statistics prepare Qumayma Al-Ahmed of the ...
The Palestinian Bureau of Statistics Central Agency for General Census of Population and Housing, the integration of the ...
Palestinian Central Bureau of Statistics Census of Population and Housing, the integration of gender Thank you for your kind ...

exc_AR.xlsx
Language: **ARA** - File name: **exc_AR.xlsx**
Acquisition: **12 Jul 2016 at 10:29 CEST** - Document date: **17 Nov 2015 at 09:26 CET**
Source name: **Test_Protocol** - Dataflow name: **20160712**
Relevant sentence for the document
Doctorat Thèses de doctorat en cours N° Nom et prénom du thésard adresse email Intitulée de la thèse Date de la première ...

Figure 7. The Tadam interface.

But we are also aware that the project takes place in the context of huge trends: Digitalization is generating an ever-growing data flood. Society is increasing in complexity. Processes are accelerating. Evidence-based (document-based) journalism has become a requirement of the profession, driven by the trust crisis in media and the fact that more and more conflicts are dragged through the courts. Finally, the professionalization of the communication units of companies, organizations and authorities, which are the institutions scrutinized in public interest journalism, is forcing journalists to work in teams.

We think that the combination of computer-assisted reporting and a system like Tadam will allow Tamedia's journalists to respond to many of these challenges with stories of greater timeliness, depth, and context.

11. ACKNOWLEDGMENTS

Our thanks to Professor James T. Hamilton from Stanford University for helpful advice.

12. REFERENCES

[1] In a recent study among more than 900 journalists in Switzerland, about 80 percent answered that "Time available for researching stories" has "decreased". - F. Dingerkus, G. Keel and V. Wyss : *Journalists in Switzerland*, ZHAW Winterthur, part of The Worlds of Journalism Study (WJS), page 5, 2016.

[2] M. Walker Guevara, deputy director of the International Consortium of Investigative Journalists (ICIJ) during a panel at John S. Knight Journalism Fellowships at Stanford, July 1, 2016.

[3] S. Cohen, J. T. Hamilton, and F. Turner : *Computational Journalism*. Communications of the Association for Computing Machinery, Vol. 54 No. 10, Pages 66-71, 2011 (10.1145/2001269.2001288).