The TRAME Project – Text and Manuscript Transmission of the Middle Ages in Europe

Emiliano Degl'Innocenti[1](✉), Alfredo Cosco[2,3], Fabrizio Butini[1], Roberta Giacomi[2], and Vinicio Serafini[2]

[1] Fondazione Ezio Franceschini, Florence, Italy
emiliano@fefonlus.it
[2] SISMEL, Florence, Italy
alfredo.cosco@gmail.com
[3] ZKS Foundation, Geneva, Switzerland

Abstract. TRAME is a research infrastructure for medieval manuscripts. The TRAME engine scans a set of sources for searched terms and retrieves links to a wide range of possible information, from simple reference, to detailed manuscript record, to full text transcriptions. Currently, it is possible to perform queries by: free-text, shelfmark, author, title, date, copyst or incipit, on more than 80 selected scholarly digital resources across EU and USA. Since 2014 September 1st, TRAME has entered a new phase and the current work is focused on: extending the meta-search approach to other web resources, leveraging the users interaction to define an ontology for medieval manuscripts, re-designing the front-end towards a new UX approach.

Keywords: Crawler · Meta-crawler · Search engine · Medieval manuscripts · Illuminated manuscripts · Digital humanities · User experience · Design · Responsive · Usability

1 Introduction

TRAME[1] was born in 2011,[2] the main aim is to build a "research infrastructure project focused on promoting interoperability among different digital resources available in the medieval digital ecosystem",[3] by connecting repositories of digitized images of medieval manuscripts, their codicological descriptions, their textual and philological interest, their cultural significance in the context of the european history. Currently it implements a number of features (including simple, shelfmark, advanced search mode etc.) on more than 80 selected scholarly digital resources around western medieval

[1] Home page: http://git-trame.fefonlus.it/.

[2] E. Degl'Innocenti, *Trame: Building a Meta Search Tool for the Study of Medieval Literary Traditions* in EVA 2011, Proceedings. Vito Cappellini, James Hemsley (eds.), Bologna, Pitagora, 2011.

[3] TRAME. *Text and Manuscript Transmission of the Middle Ages in Europe. Evolving the System Towards Horizon2020 and VCMS Challenges*. http://www.sismelfirenze.it/index.php?option=com_k2&view=item&task=download&id=68_69e648d4f36e436d0ec96c334a0180a4&Itemid=266&lang=it.

D. Calvanese et al. (Eds.): IRCDL 2015, CCIS 612, pp. 79–85, 2016.
DOI: 10.1007/978-3-319-41938-1_9

manuscripts, authors, and texts across EU and USA, including digital libraries, research databases and other projects from leading institutions.

TRAME is more than just a piece of software: it is a research tool deeply rooted in the international medieval scholarly community, whose development is in line with the Memorandum of Understanding of the COST Action IS1005 "Medieval Europe Medieval Cultures and Technological Resources", representing 260 researchers coming from 39 leading institutions (archives, libraries, universities and research centers) in 24 countries across the EU.

It has been selected for inclusion by the CENDARI e-infrastructure and is part of the DARIAH-IT landscape.

2 How TRAME Engine Works

The crawler engine is written in OO-PHP, the design follows the HMVC Pattern,[4] the RDBMS is MySql and the front-end combines Xhtml and Javascript.

Currently user can choose between more than 80 sources and perform 3 kinds of search: free-text, by shelfmark (city, library, mark) or advanced (title, author, date, incipit, copyst).

… *estendi*…

In function of the "search type", searched terms are pushed to the sources using five methods:

- **GET or POST classes**

The standard *http* method to pass variables by *query string* or a *form*, TRAME uses CURL[5] to build the request. Few sites use a minimal protocol for interoperability of medieval manuscripts.[6] Based on TEI, it provides the base set of information useful for TRAME: a shelfmark and a URL.

- **CACHED class**

Some sites are not directly searchable, others have a limited records number, some have both of these features. For those sources TRAME uses MySQL tables where it imports any possible relevant content. Tables are not public and are only used to perform a better and faster search, the result links point to the original source.

[4] "The HMVC pattern decomposes the client tier into a hierarchy of parent-child MVC layers", cf. http://www.javaworld.com/article/2076128/design-patterns/hmvc–the-layered-pattern-for-developing-strong-client-tiers.html.

[5] CURL is a library for transferring data with URL syntax, cf. http://curl.haxx.se/ and http://php.net/manual/it/book.curl.php.

[6] http://git-trame.fefonlus.it/TRAME_protocol_v1.pdf and http://www.tei-c.org/index.xml.

- **SPIRIT class**

Some sites perform searches by a javascript UI, using AJAX[7] calls to show the results. The way to query those sources is a *headless browser*[8] and TRAME uses *casperjs* and *phantomjs*[9] to do this.

Each class, customized for every source, parses the response using *reg-ex* and/or *PHP Simple Html Dom.*[10]

In closing, another class renders and composes the result as a list of shelfmarks and titles linked to the original sources.

3 Extending the Meta-Search Approach

Since September the 1st, the aim is to extend the meta-search approach to other web resources (libraries, portals, individual research projects), using various tools and technologies.

Moreover, the TRAME team is extending the engine in two other ways:

- make TRAME a tool to build a knowledge base for medieval manuscripts;
- ensure a better user experience,

3.1 New Resources

Adding new resources in TRAME implies a deep analysis on sources query methods and of the response code, until 2013 of December 12 new sources were added (Table 1).

To achieve the above we added:

1. a new library to simplify the process to identify and extract pieces of code from results along *reg-ex*:

- **Simple HTML DOM** (http://simplehtmldom.sourceforge.net/);

1. three new tools to perform searches *via* javascript and AJAX interfaces:

- **phantomjs** (http://phantomjs.org/) → An open source headless web-browser, *i.e.* the toolkit to scrape sites;

[7] http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/.

[8] A headless browser may be defined as "a piece of software, that accesses web pages but doesn't show them to any human being" (http://durandaljs.com/documentation/Making-Durandal-Apps-SEO-Crawlable.html).

[9] *phantomjs* (http://phantomjs.org/) is an open source headless browser, *casperjs* (http://casperjs.org/) is a framework built on top of it.

[10] A open source HTML DOM library, written in PHP5, that makes easy to manipulate HTML http://simplehtmldom.sourceforge.net/.

Table 1. New sites added

SITE	Class name	Type	Sources
Schoenberg Database of Manuscripts at UPENN Libraries http://dla.library.upenn.edu/dla/schoenberg	SHON	GET	220889
Manuscripts in the Library of St John's College, Cambridge http://www.joh.cam.ac.uk/library/special_collections/manuscripts/medieval_manuscripts/	SJCAM	CACHE	277
KUL List of microfilmed manuscripts http://hiw.kuleuven.be/apps/microfilm/microfilm.php	KULEUVEN	POST	4807
The MacKinney Collection of Medieval Medical Illustrations http://dc.lib.unc.edu/cdm/search/collection/mackinney/	MACKINNEY	GET	1041
Early Manuscripts at Oxford University http://image.ox.ac.uk/	OXFORDMS	CACHE	90
Beinecke Digital Collections http://brbl-dl.library.yale.edu/	BDC	GET	127641
University of Glasgow http://special.lib.gla.ac.uk/manuscripts/search/	GLASUL	POST	N/d
München http://daten.digitale-sammlungen.de	MDZ	CACHE	1300
LUND University library http://laurentius.ub.lu.se	LUND	CACHE	71
Biblioteca Municipale di Lione http://florus.bm-lyon.fr	BMLYON	CACHE	55
Univeristy Libaries in South Carolina http://digital.tcl.sc.edu/	ULSC	CACHE	264
Medieval Manuscript in Dutch Collections http://www.mmdc.nl/static/site/search/	MMDC	SPIRIT	N/d

- **casperjs** (http://casperjs.org/) → An open source navigation scripting & testing utility for PhantomJS;
- **php-casperjs** (https://github.com/alwex/php-casperjs) → An open source php wrapper for *casperjs*.

A new class, called **SpiritSite.class.php**, has been written to manage the process.

We introduced these new libraries because *CRUL* and *Simple HTML Dom* can only parse synchronous HTML and cannot interact with the page, we introduced the *headless browser* to manage asynchronous calls (AJAX) as in **Medieval Manuscript in Dutch Collections** (http://www.mmdc.nl/static/site/search/).

3.2 Building a Knowledge Base on Medieval Manuscripts

Among others, managing TRAME showed difficulty to know what users do with the site.

During the development process of the engine the issue concerning the interaction between users and the TRAME application raised, to manage that we added an *analytics* back-end, that is made by a set of php functions:

- Tracking users visiting TRAME without performing any search;
- Tracking users performing searches recording:
  - Target resources;
  - Searched terms;
- Recording the user interaction with the result sets.

During this process the team realized that those data were also useful to populate an ontology using a **bottom → up/user driven** pattern.

So we decided to start from those data to design the part of TRAME that will be connected with CENDARI[11] infrastructure following this design:

1. A user performs a search;
2. TRAME logs information about the search and builds a list with relevant resources;
3. Leveraging on above information (i.e. the log) TRAME performs the same search using a *headless browser* approach to import selected pieces of information from web pages:
   - the couple *phantomjs* + *casperjs* works fine also to build a data scraper, through which you can use CSS, Xpath or DOM selectors. Moreover the script follows relevant links to connected pages and imports again data from them;

1. The *scraper* produces a set of XML files with info about authors, manuscript, places etc. to be used by other internal or external knowledge extraction services.

The ontology is parsed by a *Name Entity Recognition* (NER), this process generates an *untrusted* ontology, it is then submitted to a validation, using specific tools, by domain experts.

The validation leads to a *trusted* ontology that can be queried by ad hoc instruments including the TRAME tool OntoQuery[12] and a SPARQL-endpoint.

Both the trusted than the untrusted ontology are hosted in a ‘triple-store’, at this time our choose was for OpenRDF SESAME[13] as front-end of Openlink Virtuoso[14].

[11] “CENDARI (Collaborative European Digital Archive Infrastructure) is a research collaboration aimed at integrating digital archives and resources for research on medieval and modern European history.” cf. http://www.cendari.eu/about-cendari/.

[12] Right now OntoQuery is developing and populated by an ontology test.

[13] “Sesame is a de-facto standard framework for processing RDF data. This includes parsing, storing, inferencing and querying over such data. It offers an easy-to-use API that can be connected to all leading RDF storage solutions.”cf. http://rdf4j.org/.

[14] cf. http://virtuoso.openlinksw.com/.

3.3 Improving the TRAME User Experience

Between 2013 and 2014 there have been some major changes in the Web, just to say two: the mobile internet exceeded the pc browsing,[15] HTML5 became an official W3C standard.[16]

So the TRAME team thought to re-design the tool, whose first step is to include the engine in a php *fast development framework*.

Why?

- Code maintenance and reuse
- Faster further code development
- Complete separation of engine core from user interfaces
- DB agnostic interfaces
- Implementation of caching mechanisms for a faster response
- Form validation
- Session handling.

Our choice was for CODEIGNITER,[17] it is an open source framework, has a small footprint, is fast and complete; moreover, there is a CMF[18] script built on top of it with some essential features. It's named NotOnlyCMS[19] and provides:

- Access Control List (ACL)
- Scaffolder
- Admin area
- HTML5 Templating with **Bootstrap**.[20]

The introduction of an ACL makes possible to add services to registered users like:

- User shelf for sets of sources
- Pre-built sources sets
- Sort results
- Export results (XML, RDF, FIRB, TEI, RSS)
- Share results on different channels (email, blog, social networks)
- Evaluate results (rating or Like).

[15] Cf. http://searchenginewatch.com/sew/opinion/2353616/mobile-now-exceeds-pc-the-biggest-shift-since-the-internet-began.

[16] To read the specs from W3c: http://www.w3.org/TR/html5/ , for an overview on creation of the HTML5 standards see Paul Ford, *The Group That Rules the Web*, The New Yorker, Nov 20, 2014, http://www.newyorker.com/tech/elements/group-rules-web.

[17] CODEIGNITER (http://www.codeigniter.com/) is maintained by the British Columbia Technology Institute (http://www.bcit.ca/cas/computing/).

[18] CONTENT MANAGEMENT FRAMEWORK, a Framework with common pre-built CMS-like features.

[19] The code is on Github: https://github.com/goFrendiAsgard/No-CMS.

[20] Bootstrap is a HTML5 framework built by *twitter* and released free, cf. http://getbootstrap.com/.

The Bootstrap integration means that the next UI will be **prototyped** and **responsive;** furthermore, to get a better **usability** the implementation of a Bootstrap extension called **Assets Framework** is on the ground:

> *"Assets gives you Sect. 508 compliant, cross-browser compatible UI components that you can use in your accessible web site or web application. Assets is an accessible, responsive, and modern framework."*[21]

3.4 TRAME Goes Social

Renewing the project we decided to pay attention to the Internet and Social Networks, a blog has been activated, that will start on January/February 2015:

http://trameproject.blogspot.it/.

We release news in two SN channels too:

Twitter: https://twitter.com/trameproject

Facebook: https://www.facebook.com/trameproject

[21] *cf.* http://assets.cms.gov/resources/framework/3.0/Pages/, *for further information on Sect. 508 see:* http://www.hhs.gov/web/508