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

Emiliano Degl'Innocenti*, Alfredo Cosco**, Fabrizio Butini*, Roberta Giacomi***, Vinicio Serafini***

*Fondazione Ezio Franceschini

emiliano@fefonlus.it,

**SISMEL, Firenze, Italy / ZKS Foundation, Genève, Switzerland

alfredo.cosco@gmail.com

***SISMEL Firenze

Abstract. TRAME is a research infrastructure project focused on medieval manuscripts, authors and texts. The TRAME engine scans a set of sources for searched terms and retrieves links to a wide range of possible information (i.e.: simple references, detailed manuscript records, full text transcriptions, etc.). Currently TRAME allows users to search using both simple (i.e.: free-text) and advanced (i.e.: search for: shelfmark, author, title, date, copyst and incipit) methods on more than 80 selected scholarly digital resources across EU and USA. Recently (September 2014) the TRAME development has entered in a new phase 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 better UX.

Keywords: crawler, meta-crawler, search engine, medieval manuscripts, illuminated manuscripts, digital humanities, user experience, design, responsive, usability, interoperability, knowledge extraction

1. Introduction

TRAME[1] was born in 2011,[2] the main aim was 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

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.

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 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[4] e-infrastructure and is part of the DARIAH ERIC landscape.

2. Methodology

The first release of TRAME was aimed at improving the level of interoperability of digital resources and datasets available for medieval researchers and scholars, provided by archives, libraries and research centers. The main goal was to fill the gap between the researchers' needs (i.e.: to found valuable information related to medieval manuscripts, authors and texts) and the fragmentation and lack of interoperability in the medieval digital ecosystem (i.e.: between digital resources made available by libraries and archives).

The starting phase (2011) was focused on the meta-search approach in order to build a simple and thin tool to dispatch the users' queries on a number of selected sources, using just on-the-fly searches, with no data collection or query extension mechanisms. The first release of TRAME was published in 2011, then the project joined two European networks: COST and CENDARI.

Since then the TRAME development agenda was focused on building a Medieval Semantic Knowledge base, by developing tools for information collection (i.e.: web crawling and data mining) and semantic integration, in order to allow users manage complex research questions instead of performing traditional queries. The design and implementation process of the TRAME crawler is the main focus of this paper.

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.

4 "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/.

3. How TRAME engine works

The crawler engine is written in OO-PHP, the design follows the HMVC Pattern,[5] 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).

Tab.1 Detail of the TRAME search fields

Name	Description	Notes
freetext	Freetext search on all fields	A boolean operator is used to the remote website logic
author	Search for specific medieval author(s)	Values are coming from domain Thesaurus developed in the context of IS 1005 COST Action
title	Search for specific work title(s)	
incipit	Search for work(s) incipit	
date	Search for manuscript datation	User could search using descriptive labels (e.g.: XIII c.) or numeric values (e.g.: 1201-1300)
location	Search for a specific manuscript	User could search using the extended shelfmark of a single manuscript (e.g.: City, Library, Collection) using a shared authority list developed in the context of IS 1005 COST Action
shelfmark	Search for a specific manuscript	User could search using the numeric shelfmark of a single manuscript (e.g.: vat. Lat. 3195) using a shared authority list developed in the context of IS 1005 COST Action
copyist	Search for specific medieval copyst(s)	

According to the selected "search type", searched terms are pushed to the sources using five methods:

- **GET or POST classes**

The standard *http* methods to pass variables by *query string* or a *form*, TRAME uses CURL[6] to build the request. Few sites use a minimal protocol for interoperability of medieval manuscripts.[7] 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 resources TRAME uses tables in a local

5 "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.

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

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

MySQL DB to store any possible relevant content. TRAME uses the cached information to manage the queries, allowing the results to be accessed directly on the original provider website according to the their access policies.

- **SPIRIT class**

Some sites perform searches by a javascript UI, using AJAX[8] calls to show the results. The way to query those sources is a *headless browser*[9] and TRAME uses *casperjs* and *phantomjs*[10] to do this.

Each class, customized for every source, parses the response using *reg-ex* and/or *PHP Simple Html Dom.*[11]

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

4. 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,

4.1 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.

Tab.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	MACKINNEY	GET	1041

8 http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications/.

9 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).

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

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

http://dc.lib.unc.edu/cdm/search/collection/mackinney/			
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
French Illuminated Manuscripts http://www.enluminures.culture.fr/documentation/enlumine/fr/	ENLUMINURES	CACHE	1740
Medieval Manuscript in Dutch Collections http://www.mmdc.nl/static/site/search/	MMDC	SPIRIT	N/d

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/);
2. 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;
    - **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/).

4.2 Building a Knowledge Base on Medieval Manuscripts

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 the CENDARI infrastructure following this design:

1. A user performs a search;
2. TRAME logs information about the search and builds a list with resources (i.e.: site title and URI);
3. Leveraging on the above information (i.e. the log) TRAME performs the same search using a *headless browser* component to import selected pieces of information from web pages using *phantomjs+casperjs*. Moreover, a rule based data scraper component (using CSS, Xpath or DOM selectors) has been created to follow relevant links to related pages and collect information;
4. 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 result is parsed by a *Name Entity Recognition* (NER) tool, in order to provide candidates for inclusion in the Medieval Semantic Knowledge base, after a validation process led by domain experts.

The results could be queried by ad-hoc instruments including the TRAME tool OntoQuery[12] and a SPARQL-endpoint.

Currently the knowledge base is hosted in a Openlink Virtuoso[13] triple-store, using OpenRDF SESAME[14] as front-end.

4.3 Improving the TRAME User Experience

During the last couple of years a number of changes hit the WWW, just to say two: the mobile internet exceeded the PC browsing,[15] HTML5 became an official W3C standard.[16] TRAME was then re-designed to wrap the engine in a php *fast development framework*.

12 Right now OntoQuery is developing and populated by an ontology test.
13 cf. http://virtuoso.openlinksw.com/
14 "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/
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.

In order to seek faster code development, maintenance and reuse, allow performances improvement, a complete isolation of the engine core from the user interfaces, implement advanced caching mechanisms and add form validation and session handling features we used CODEIGNITER,[17] along with NotOnlyCMS[18]. The former is an open source framework, with a small footprint, fast and complete; the latter is a CMF[19] script built on top of it, adding some nice features such as: Access Control List (ACL) management, scaffolder and admin area, HTML5 Templating with **Bootstrap**.[20]

In particular, the introduction of an ACL management system makes possible to add custom 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).

Furthermore the Bootstrap integration allows for the UI to be **prototyped** and **responsive**, in order to reach a better **usability**; the implementation of a Bootstrap extension called **Assets Framework** is in progress:

> *"Assets gives you Section 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]

4.4 TRAME dissemination and outreach

Users could get updated information about the development of the TRAME project through the dedicated blog[22] and the related social network channels, on Twitter[23] and Facebook[24]

5. Conclusions and next steps

TRAME is an ongoing collaborative international effort, rooted in the medieval research community. Its development agenda is deeply influenced by the needs expressed by scholars across EU and US. Recent changes about the nature of the information available

17 CODEIGNITER (http://www.codeigniter.com/) is maintained by the British Columbia Technology Institute (http://www.bcit.ca/cas/computing/).
18 The code is on Github: https://github.com/goFrendiAsgard/No-CMS.
19 CONTENT MANAGEMENT FRAMEWORK, a Framework with common pre-built CMS-like features.
20 Bootstrap is a HTML5 framework built by *twitter* and released free, cf. http://getbootstrap.com/
21 cf. http://assets.cms.gov/resources/framework/3.0/Pages/, for further information on *Section 508* see: http://www.hhs.gov/web/508.
22 cf. http://trameproject.blogspot.it
23 cf. https://twitter.com/trameproject
24 cf. https://www.facebook.com/trameproject

in the WWW influenced the development of TRAME from a mere meta search approach towards the establishment of a Medieval Semantic Knowledge base, using custom modules for information collection and integration (i.e.: web crawler, data miner) as described in this paper. The new release of TRAME – with the described improvements – will be tested and published in the second half of 2015.