Documenting some Uses of the Isidore Platform

Sophie David, Jean-Luc Minel, Stéphane Pouyllau

Documenting some Uses of the Isidore Platform

Sophie David (TGE Adonis, CNRS)
Jean-Luc Minel (Modyco, Université Paris Ouest Nanterre La Défense et TGE Adonis)
Stéphane Pouyllau (TGE Adonis, CNRS)

Abstract
The availability of millions of digital resources, and of the means for processing, analysing, signalling, and exposing them, should make it possible to facilitate the emergence of new research objects and new questions. After a brief description of Isidore, this paper proposes first results in observing the Isidore platform because it presents all the basic materials a researcher needs: publications, blogs, websites and conference calendars. Then we discuss how facets proposed by Isidore are currently being used and finally we propose some issues which are being studied.

Introduction
It is increasingly recognized that Digital Humanities will change researchers' practices (ThatCamp 2010, Juanals & Noyer 2010, Vinck 1991). The availability of millions of digital resources, and of the means for processing, analysing, signalling, and exposing them, should make it possible to deal with old problems in new ways, or facilitate the emergence of new research objects and new questions, in particular because of interdisciplinary exchange and the creation of novel web-enabled links, etc. which were initially inconceivable (this is for instance the position of the Dariah project). Nevertheless, it remains difficult to precisely objectivise these expected benefits. It is true that it is still early days. It is also true that only a few surveys have documented and analysed these new or innovative usages (Garron & Minel 2010, Gallezo & Noyer 2010).
Documenting and analysing usages requires setting up methods of observation (Le Marec 2001). Several are possible, depending partially on the chosen "object", partially on the observation framework (evaluation, documentation, etc.), and above all, on how the objects to be evaluated are considered (Amar *et al*. 2008).
We are interested in observing the Isidore system (http://www.rechercheisidore.fr/apropos), because it presents, on a single platform, all the basic materials a researcher needs: publications (to determine the state of the art, to access documents, etc.); blogs (to follow research in the making, to track research developments, etc.); websites (to access results, to belong to a community, etc.); and conference calendars (to monitor the state of the art, to announce results, etc.).
We propose to analyse different uses of the platform, basing the analysis on elements provided by Google Analytics. Three immediate interests concerning Isidore itself will be addressed: a) to provide some initial answers to a set of precise questions about the use of Isidore; b) to identify the elements which could lead to improvements in the platform; c) to identify the elements which could improve observation of how the platform is used. There is a fourth, additional interest: from the collected documentation, to begin to develop the problematics and methods to observe "composites"[1] (Le Marec 2001), i.e. to carry out a survey of the usage of Isidore. We will first give the key elements for understanding the operating model of Isidore. Second, we formulate the sets of questions we seek to answer; we then briefly describe the corpus of statistics on which our analyses are based. Lastly, we present our results and analyses.

1. What is Isidore?

[1] Composites means "work situations, in which people mobilize material objects and representations […], carry out actions […], and operate norm systems or operating rules" (Le Marec 2001: 187). Field surveys are therefore favored.

The Isidore platform (Isidore 2011) is both a technical search engine and a new approach to designing a flexible workstation for targeted end-users, namely researchers, students, engineers, and librarians working in the Humanities. Isidore is based on radically different principles to Google, Bing, etc. First, the crawled sources were carefully chosen after a preliminary study of end-users' needs. Consequently, sources come from different categories of web pages: journals, blogs, RSS feeds and announcements of scientific events. Secondly, a panel of thesauri and list of references (produced by the scientific community) were selected in order to enhance the indexing process. Thirdly, metadata are the cornerstone of the whole process, from crawling to the user interfaces. Isidore is powered by different modules which exchange information, metadata and data, connected by the web of data and linked data methods, and stored in an RDF triple-store. Isidore, as a semantic web system, also offers a SPARQL endpoint.
With regard to the question of user interfaces, results are displayed without any ranking order but several selection filters (or facets), around 10, are available. The choice to use facets rather others query refinement systems is based on several studies (Yee *et al.* 2003, Dakka *et al.* 2008) which have shown how they upgraded the relevance of the query. These facets enable the end-user to navigate through the results in order to find relevant sources. Furthermore, by analysing results and data linking (Isidore uses linked data methods) some suggestions are put forward.

2. Questions, Results and Analyses

Our aim is to answer four sets of questions (only three of which will be tackled here[2]):
- Do users formulate the different types of queries which are available?
- Do users use the facet system? To what extent? What is the distribution of facets?
- Do users browse on thanks to the enrichment provided by hints? To what extent? Do they use generalist thesauri or specialized ones?

The statistics used here come from Google Analytics. The period observed is the third week in July 2011[3]. We have analysed 2385 queries, which represents 3377 pages and 2702 unique visits. In quantitative terms, that represents a "weak" week of activity (when compared with the Piwik results).

2.1. Four Types of Use

Four specific uses can be identified:

	Direct query	More options query	Facet exploration	Hints query	Total
Number	922	732	372	359	2385
%	38.7	30.7	15.6	15.1	100

By 'direct query' (a) we mean a query where the user writes an expression in the dedicated box on the home page; by 'more options query' (b), a query where the user writes the query using the more options possibilities (one or several options): author, title, term, before, after; by 'facet exploration' (c), a query where the user explores the resources with the facets but does not write any expression; by 'hints query' (d), a query where the user, after a query of type (a) or (b), uses the enrichment to continue browsing.
Almost 70% of users write expressions, which is not very surprising. Queries of type (c) and (d) account for more than 30%, equally divided, and are therefore significant from a quantitative point of view. It thus seems relevant to examine the facet system and the use of enrichment in greater detail.

[2] It seems relevant to us to analyze the form of the queries from a linguistic point of view (simple terms *versus* complex terms, expressions, sentences, grammatical categories, proper names, etc.). We can then assess the possible influence of a Google search. This point is important, because it outlines a specific appropriation by users, and determines possible corrective actions, additional explanations, etc. We have used Piwik, an open source service, which was put in place on the beta version of Isidore; we have statistics for more than 6 months of use. Results will be given in the final version, if this paper is accepted.
[3] In the final version, we will extend the period in order to get statistics on several weeks and to check the representativeness of our results.

2.2. Using Facets

Facets are used to explore the resources (cf. the facet exploration). But they can also be used to restrict or to shift a query scope. Overall, 26.2% of the queries are constrained by one or several facets (i.e. 630 queries out of 2385, which represents 941 facets; up to 5 different facets can be used in the same query). It is also interesting to cross-link these constraints with regard to the types of queries identified above:

	Direct query	More options query	Facet exploration	Hints query
Constrained by at least 1 facet	341	34	232	23
% (number of queries = 2385)	14.3	1.43	9.73	0.96
% (number of queries in each query type)	36.98	4.64	62.37	6.41

Very strong contrasts can be observed: facets are used a) because the query is not at first contextualized (cf. direct query versus more options query, in which users can type more items of information (author, title, etc.)), or b) because exploration has been chosen (facet exploration). A little more surprising is the low rate of hints queries (which should resemble the direct query rate, because a hints query is like a new query). No doubt the fact that the hints query is not immediately available, requiring formulation of a query, access to one specific resource, access to this service (which presupposes that the need is real and the service is well understood), could explain this result. However, another explanation (apart from the hypothesis of a bias, which is also possible) can also be given. The hints query is used as another way to restrict the number of results; if the meaning of the chosen term is sufficiently specified in the context, there would be no need to restrict the search even further[4]. The distribution of the facets is the following:

	Type of documents	Discipline	Century	Language	Historical period	Categories	Collection and Institution	Others	Total
Number	270	214	185	78	64	50	54	26	941
%	28.69	22.74	19.66	8.29	6,80	5,31	5,74	2.76	100

2.3. Enrichment

Three thesauri are used: Rameau, Pactols and Geonames. At the moment, users can use hints query only with Rameau and Pactols. The results are displayed below:

	Rameau	Pactols	Total queries[5]
Number	345	15	359

It can be seen that there is a great difference in the use of the two thesauri. Two explanations can be suggested: the version of Pactols used in Isidore contains approximately 60,000 terms, while Rameau contains approximately 200,000 terms. Rameau is much more heavily used simply because it contains far more terms. Second, Rameau is a generic thesaurus, while Pactols is archaeology- and history-oriented.

Conclusion

These different observations are the first step in the analysis of the use of Isidore. We can confirm that the four types of queries are used, as well as the facet system and the hints system. However, it appears that having access to the 'routes' chosen by the users would be a fruitful way to improve our analyses: we could produce a better analysis of the enrichment; we could have a better idea of

[4] Furthermore the user knows in advance how many resources are tagged with this term. Of course, with this method of observation, we can not know how many trials have been done before a click.

[5] One user formulated a query using both thesauri. That is why the number of queries is 359 and not 360.

other uses of the facet system or the hints system, which should perhaps not be analysed only as restrictions. It would also enable us to answer other types of questions: are there exclusive uses of Isidore? For example are there users who are just interested in calendars, just in blogs, or just in PhDs? Or can types of query chaining be identified? In any case, we need better tools to answer more sophisticated questions.

References

Amar, M., David, S., Panckhurst, R., Whistlecroft, L. 2008. Classification Procedures for Software Evaluation. Actes du colloque LREC, Marrakech, mai 2008.

Dakka, W., Ipeirotis, P. G., Wood K. R. 2008. Automatic Extraction of Useful Facet Hierarchies from Text Database, ICDE

Gallezo, G., Noyer, J.-M. 2010. De la numérisation des revues à l'expérimentation d'une édition de recherche processuelle, in Technologies de l'information et intelligences collectives, Juanals, B., Noyer, J.-M. (ed.), Lavoisier, Paris.

Garron, I., Minel, J.-L. 2010. Formes de lecture et usages du Web, in Technologies de l'information et intelligences collectives, Juanals, B., Noyer, J.-M. (ed.), Lavoisier, Paris.

Isidore 2011, http://www.rechercheisidore.fr/apropos.

Juanals, B., Noyer, J.-M. 2010. De l'émergence de nouvelles technologies intellectuelles, in Technologies de l'information et intelligences collectives, Juanals, B., Noyer, J.-M. (ed.), Lavoisier, Paris.

Le Marec, J. 2001. Ce que le « terrain » fait aux concepts. Vers une théorie des composites. Thèse d'habilitation. Université Paris 7.

Manifesto for the Digital Humanities, THATCamp Paris 2010, http://tcp.hypotheses.org/411, [d.c. 2011-07-26].

Vinck, D. 1991. La gestion de la recherche. Nouveaux problèmes, nouveaux outils, De Boeck, Bruxelles.

Yee, K., Swearingen, K., Li, K., Hearst, M. 2003. Faceted metadata for image search and browsing, Conference on Human Factors in Computing Systems.