How to cite:

Rodríguez Ortega, Nuria and Cruces Rodríguez, Antonio. «Development of Technological Ecosystems for Cultural Analysis: The Case of Expofinder System and Art Exhibitions», *Digital Scholarship in the Humanities* (Oxford University Press), 34, n.º 2, 2019, pp. 423-448. ISSN: 2055-768X (online) (first published: August 2018) (postprint. doi: **10.5281/zenodo.10446598**)

Development of Technological Ecosystems for Cultural Analysis: The Case of Expofinder System and Art Exhibitions

Nuria Rodríguez-Ortega (UMA) and Antonio Cruces Rodríguez (UMA)

Abstract

Cultural analysts are currently faced with an ecosystem of heterogeneous, globally distributed, and mass-scale sources that cannot be ignored in the research process. In addition to that, cultural analysts need to deal with the emergence of a knowledge economy which is no longer based on the value of the information entities contained in one or more documents, but on the potential for heterogeneous data to be recombined to generate previously unknown knowledge. Therefore, the challenge of accessing a set of heterogeneous, mass-scale, dynamic, and globally distributed sources has been joined to that of transforming the content of these sources into reusable data for the creation of knowledge and value. To the traditional question—What sources should be consulted or selected?—must now be added: What tools should be built and what work procedures should be designed to access those sources on a mass basis and analyze them as data? This article aims to provide some possible answers through the work carried out within the Exhibitium Project. This included the design and implementation of the Expofinder system, which is a technological device meant to examine the mechanisms for dissemination of digital information about art exhibitions and reuse this information as data to generate new knowledge and new interpretations about them. We believe that one of the best ways to strengthen the paths of digital research is to make the methodologies and mechanisms that govern decision-making transparent in order to be discussed or to be adopted by other projects. This is the ultimate goal of this article.

1. Cultural–Historical Sources in a New Scenario

The statement that the changes brought about by the digital society entail disciplinary changes related to methodological processes and research practices is already customary in the scholarly digital discourse. Sources, understood as the contexts from which relevant data for cultural research are extracted, are directly involved in these changes due to the unprecedented increase in the number of documents and information currently accessible. This has largely resulted from the proliferation of digitization projects and the building of digital archives and repositories, as well as from the huge volume of content that is directly produced in the digital space on a daily basis. Therefore, in the economy of abundance, the seemingly unlimited access

and the exhaustiveness of the sources are the strands from which cultural analysts must now weave the fabric of their primary materials (Rodríguez-Ortega, 2015).

It also needs to be considered that exhaustiveness does not only refer to the vast amount of data, records, documents, and information that can be found on the numerous web devices currently avail- able. Exhaustiveness also involves two other aspects. These are the exponential multiplication of media, materials, and contexts capable of functioning as 'sources' for cultural research, which lie beyond the usual repositories and digital platforms (archives, databases, documentary platforms, etc.)[1] and so exceed any previous typological classification as well as the global distribution of sources that can be potentially relevant to research. Moreover, these sources are also dynamic by definition since they are subject to an incessant process of change and transformation. Therefore, contemporary cultural analysts must deal with an ecosystem of heterogeneous, mass-scale, dynamic, and globally distributed sources that cannot, or at least, should not, be ignored.

This scenario creates an ambivalent situation, which involves, on the one hand, the need to expand the traditional repertoire of sources and materials and, on the other hand, the need to establish parameters and guidelines to assist researchers in dealing with this information overflow. Researchers are then faced with a certain feeling of crisis and conflict. Since the mere possibility of having access to seemingly limitless and exhaustive sources confronts researchers with the responsibility of taking them into account and relocating their work in a context of global analysis, which Fernando Marías (2008) had already pointed out in the past, the inability to cover everything emerges as a supposedly insurmountable barrier. It is no wonder, then, that one of the basic aspects of the problem has focused on the need for filters, quality criteria, and critical reviews of digital sources to select those that are the most relevant and 'authorized'.

It should be remembered, however, that the problem of selecting from large sets of sources is not unique to the digital society. It is assumed that, as we approach contemporary times, the corpus and types of sources grow and expand. To address this situation, the usual practice is well known to all. This includes providing a precise definition of the subject and limiting the scope of the field of study as a basis for developing a strategy for the coherent and meaningful selection of sources within that domain. This methodology, without apparent changes, is the same that is applied in the digital realm. In this way, the principle of a 'significant and consistent selection' still prevails in either analogue or digital forms.

These lines of action are certainly fundamental to our digital context. However, this approach does not involve a major difference from the traditional processes of academic research since the 'selection' of sources and the delimiting of the documentary field to be analyzed is an idiosyncratic practice in Humanities research cultures. The change lies here only in the amount. The selection must now be performed on a broader set of sources, but basically, the problem is still one of selecting and setting limits.

The situation varies when one takes the perspective of the macroscopic analyses linked to the—somewhat controversial—distant reading of Franco Moretti (2013) or the cultural analytics studies carried out, for example, by Lev Manovich (2015). These perspectives precisely emphasize the value of exhaustiveness and 'massiveness' as defining factors of the new epistemic conditions of our contemporaneity as opposed to filtering and representative selection.

The macroscopic perspective redefines our questions regarding sources and, consequently, initial methodological presuppositions. The issue here is not to determine what criteria make a set of sources meaningful to explain certain cultural–historical facts but how to gain access to the exhaustive set of sources that would make it possible to explore the significance of such a domain in a comprehensive manner[2]. Logically, the idea of absolute exhaustiveness is only possible as a Kantian intellectual proposal since even in macroscopic analyses there is a certain selection of the sources involved. The notion of exhaustiveness, therefore, has to be understood as a methodological principle, which in itself implies a redefinition of the positioning of cultural analysts in relation to sources.

Another circumstance is to be added to the characterization of this new scenario, including the move from the web of documents to the web of data. As a consequence, a knowledge conception emerges that is no longer based on the value of the issues of information entities contained in one or more documents and their hermeneutic analysis, but on the potential for heterogeneous data to be recombined through algorithms to generate previously unknown knowledge (descriptive and/or predictive). In other words, researchers have to place themselves in a con- text that moves from the information society to the knowledge society in which an 'explosion' of the traditional concept of sources is taking place. Rather than, in the information contained in a series of documents, the interest now lies in the data that can be extracted and relocated in a continuum of data where they can be interconnected in sometimes unexpected processes for the production of meaning.

Consequently, the challenge of accessing a set of heterogeneous, mass-scale, dynamic, and globally distributed sources has been joined to that of transforming the content of these sources into reusable and recombinable data for the creation of knowledge and value. It is impossible to meet these challenges without integrating computer technology as a constituent part of any methodological approach. This does not mean 'using' technological devices or incorporating digital platforms and archives into our traditional repertoire of sources, but 'building' technological tools and developing procedures that can provide answers to the problems posed in each case. Therefore, to the traditional question: What sources should be consulted or selected? must now be added: What tools should be built and what work procedures should be designed for their access, use, and exploitation on a mass scale? Researchers, therefore, as claimed by Alejandro Piscitelli (2015), are becoming aware that their status is no longer that of mere 'publishers' but of 'makers' as well.

2. The Exhibitium Project as a Case Study

The origin of the Exhibitium Project[3] discussed in this article was born out of the need to find answers to this new scenario faced by cultural–historical research. Specifically, the question this project set out to answer was how to take advantage of the universe of information related to cultural activity that is distributed and scattered throughout the Internet space to generate new knowledge and value. To think accurately about a work prototype, it was decided to take as a central object of study the art exhibitions regularly held in galleries, museums, and art centers since they have become a major focus for a variety of fields, including art history studies, art market dynamics, leisure industries, museum agendas, and more. Exhibitions are not only one of the main legitimizing discourses of the art system but also strategic factors that generate sociopolitical and economic movements. In addition to that, they generate a broad and heterogeneous set of digital information. Indeed, the Exhibitium Project was first determined by choosing the contemporary sources from which to extract information about art exhibitions. Although it is true that digital sources for art exhibitions coincide, to a large extent, with the institutions that organize and/or fund them (museums, art centers, galleries, universities, foundations, cultural interpretation centers, exhibition halls, and art fairs), given that they themselves act as information distributors through their dissemination channels, the information sources are broader than that and include all those web spaces which could potentially generate information about art exhibitions. This may entail personal and/or individual blogs devoted to cultural communication and contemporary critique, cultural institutions' blogs, digital magazines, and more. Therefore, art exhibitions perfectly meet the characteristics of the new types of digital sources (multiple, mass scale, heterogeneous, distributed, and in continuous flux).

However, the first thing we could observe was the almost total absence of available structured data sets or repositories. In Spain, for example, they practically do not exist. Yet, in those contexts where we find a more data centric awareness, that is, practices oriented to the production of data in open and structured systems according to standard description models, art exhibitions have not been the focus of attention for the moment. Institutions such as the Tate Galleries[4], the British Museum[5],

the Rijksmuseum[6], the Smithsonian American Art Museum[7], and the Metropolitan Museum of Art[8] have published their data sets of collections and artists but not the exhibitions' ones. As far as we have been able to know, only the Cooper-Hewitt National Museum of Design[9] has published its data set about exhibitions, although the data model is quite limited. Only a few months ago, the Museum of Modern Art (MOMA) proceeded to publish its complete exhibition data set with an interesting granularity[10]. On the other hand, semantic web strategies developed so far are limited as regards specifically art exhibitions. It is almost impossible to find data sets in the Linked Open Data (LOD) cloud that allow us to reuse them by com- puter applications or through Protocol and RDF Query Language (SPARQL) queries. Moreover, the use of specific meta tags in the code of the web pages that could help to find and extract information is anecdotal, and when they exist, their granularity is minimal. Added to this situation is the fact that a large part of the digital information on art exhibitions appear semi-structured or completely unstructured and the structures, when they exist, which do not always happen, are heterogeneous and variable[11]. This diffused and unstable landscape is also a symptom of the great diversity and heterogeneity that define the processes of description and documentation of art exhibitions and provide evidence of the lack of consensus about the type of information that should be recorded and preserved[12].

Summarizing, what we found was a complex scenario defined by a multiplicity of distributed sources, providing heterogeneous information with different levels of structuring and without previously available structured data sets and/or codified in terms of the semantic web. This finding was a determining factor in the design of the Expofinder system, which is the prototype developed within the framework of the Exhibitium Project to carry out the task of finding and capturing digital information about art exhibitions for subsequent compilation in the form of structured and interrelated data.

2.1 *Expofinder: Design and development*

In view of this approach, our initial challenge was to create a technological device capable of processing the diversity of sources available on the Web susceptible of being a vehicle for information on art exhibitions, as well as producing a platform to compile and structure the information conveyed by these sources for later exploitation and analysis. This is the Expofinder system, which is a prototype available at www.expofinder.es.

Accordingly, the first phase of the Exhibitium Project—as discussed here—was dedicated to solving three research problems, which are outlined below:

(a) The building of a device for capturing information on art exhibitions from any web source.
(b) The development of a model for recording this information in the form of structured and interrelated data for computer processing.
(c) The implementation of a workflow to guarantee the systematic recording of data as well as their quality.

2.1.1 *The WordPress framework*

With regard to developing the software for the platform, after some preliminary versions based on our own implementations, it was decided that the most interesting option out of the existing solutions, based on free and open-source software, would be to use WordPress (WP) as a framework for the system. We consider that the use of WP as a development framework represents one of the most relevant aspects put into play by the Exhibitium Project (Cruces Rodríguez *et al.*, 2016), since this demonstrates how complex systems can be built bringing to their maximum capabilities the functionalities provided by WP. Why did we decide to use WP instead of other open-source frameworks? The reason that sustains this decision is eminently technical. Therefore, unlike other specialized software such as Omeka, absolutely all WP internal operations are available through an API based on hooks that allow us to intervene and modify their actions before initiating them, during the process, and later on. In addition to that, the management of charts and maps for data visualization becomes easy by incorporating routines in the Expofinder code based on the use of open-access libraries such as

Table 1 WP subsystems used in the Expofinder system

WP subsystem	Expofinder functionalities
Ajax	Asynchronous functions executed using client-side resources
cURL	Automatic remote data capture. Read as 'see URL'
Database	DBMS direct access. Maintenance
Mail	Communication channel with users. Security
REST	Data export. Ajax (functions executed using client-side resources)
RSS/Atom Reader	Automatic remote data capture
Sanitization	Intrusion security
Users	Granular access to Expofinder functions. Session control
WP Cookies	Access control. Session control
WP Cron Jobs	Automatic remote data capture. Maintenance

Leaflet or Google Maps for SIG, D3.js, or Google Charts for graphics. The implementation built on WP also allows the exportation and reusability of data, which is explained in Section 2.1.5. In brief, the main benefits that the use of WP as framework offered to our project can be synthesized in the following items: a database with a flexible and solid organizational structure, a layer of a core application with numerous hooks which allow to maximize its functionality, and an easy management system to carry out tasks on both sides (server and client). Therefore, for implementing the Expofinder system, we took advantages of a predefined database, the available APIs, and the set of data visualization templates. The following matrix (Table 1) shows the WP subsystems used to build the Expofinder system.

To better understand the contribution of Expofinder from the point of view of the use of WP, it may be interesting to cite two previous pro- jects as examples. It must be said that, playing with this structure, at least two major projects have al- ready been developed with the purpose of trans- forming it into a universal model: WP MVC[13] by Tom Benner, and Themosis[14], by the homonymous Belgian agency. However, while these two projects are based on the implementation of an MVC mechanism (Model, View, and Controller), which implies adaptations and modifications on the WP core, the Expofinder system has been built using WP without adaptations. This means it can be downloaded from the Internet. We believe that these two examples illustrate the difference between Expofinder and other developments based on WP and allow to better understand the 'novelty' introduced by Expofinder, which is precise in its simplicity. The two projects cited are examples of implementing a general methodology for using the WP API through a second level of API on which the final development is built. This involves adding layers and inter- mediate processing and, consequently, greater complexity. In contrast, Expofinder makes direct use of the first level of API, which are the WP's ones. All the functionality of our application lies, then, on the WP code itself. Therefore, it is not supported on variant versions (forks) of the original software.

From the beginning, it was decided that the programming work should not constitute a 'tailored suit' for the Exhibitium Project. Conversely, we expect that our work can be useful in other projects (Section 2.1.5). For that reason, our choice was to implement Expofinder by means of a WP theme, which is a solution that easily allows us to readapt the software for different purposes to the extent that the entire logic of Expofinder is managed with a unique set of scripts. This is also what guarantees that the current structure of the system is capable of surviving the numerous and frequent changes that the evolution of WP undergoes. The system is de- signed as a monolithic block with all the logic contained in a WP theme. It is a folder installed in/wp-content/themes that interacts with the WP API but without modifying it. Expofinder benefits from every WP update whose successive versions can be installed without affecting its operation at all.

With regard to the extraction of Web data, we followed Baumgartner *et al.* (2009) who split this task into five different functions, including (1) web interaction, mainly comprising navigation to predetermined destination pages that usually contained the sought information, (2) extraction of data using a program that identifies them on the processed pages and then extracts and transforms them into a structured format, (3) a timeline to perform these tasks repeatedly in an automated

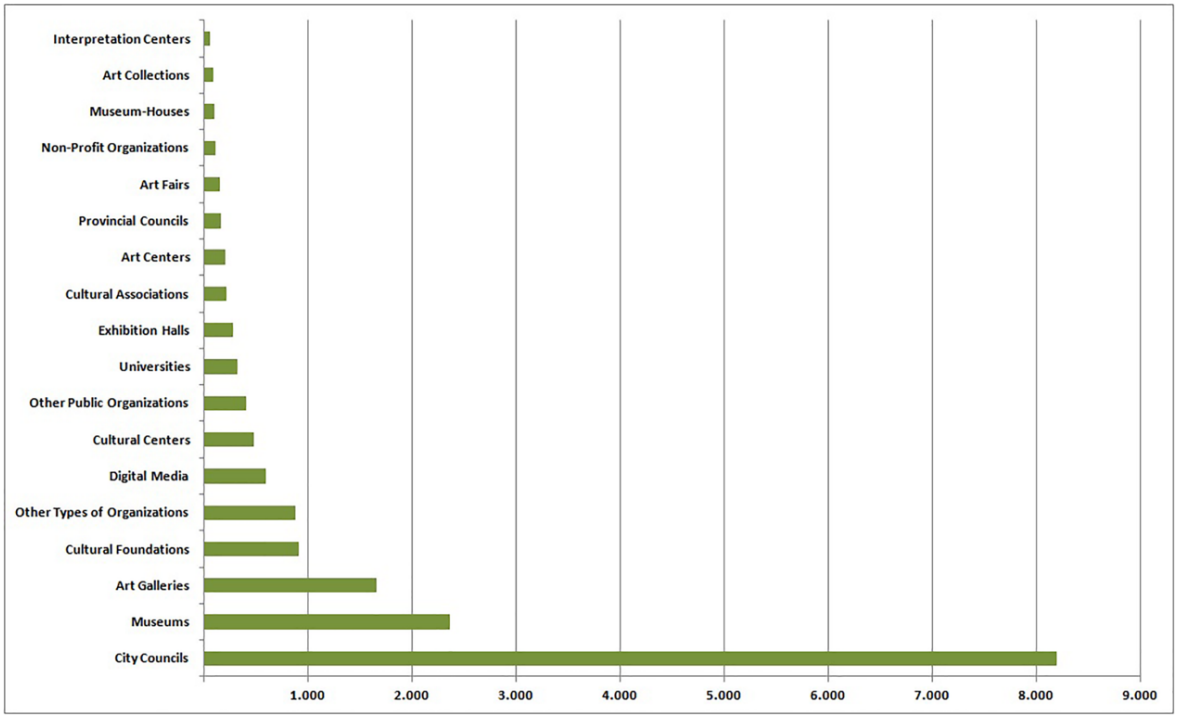

Fig. 1 URIs' distribution by source type
Source: Expofinder

way, (4) processing of the captured data, which includes filtering, refining, and integrating processes, and (5) delivery of the resulting structured data to the applications responsible for their analysis. This distribution of tasks was grouped into two large blocks within the overall Expofinder system.

(a) An information capture system as automated as possible, which would reasonably ensure the reliability of the information collected. This block is called Beagle.
(b) The entire set of elements required to store the captured information in the form of data, including filtering, cleaning, manage- ment, structuring, and description functional- ities. This block is called ExpofinderDB.

2.1.2 *Capturing and filtering dynamic information about art exhibitions*

The design and building of the Beagle + ExpofinderDB system represents one of the most relevant aspects of the Exhibitium Project. That is why we will extend longer in this section. As indicated previously, Beagle automatically captures information about art exhibitions from any web source. The web sources are recorded in the Expofinder system by their Uniform Resource Identifier (URIs). Two types of URIs are considered, which include HTML URIs and RSS/ATOM URIs. Currently, the Expofinder system comprises more than 16.000 URIs that generate an average of 50 daily news items about art exhibitions. Fig. 1 displays the number of recorded URIs and their distribution by source type[15].

In general terms, the procedure designed to find information through the URI scans consists of two steps, which are establishing a process of daily automated, repeated capture by Beagle of any new information generated on the Internet about art exhibitions from the web sources recorded in Expofinder, and implementing a filtering mechanism to refine the capture process based on the relevance of the content regarding the subject area under examination (Fig. 2).

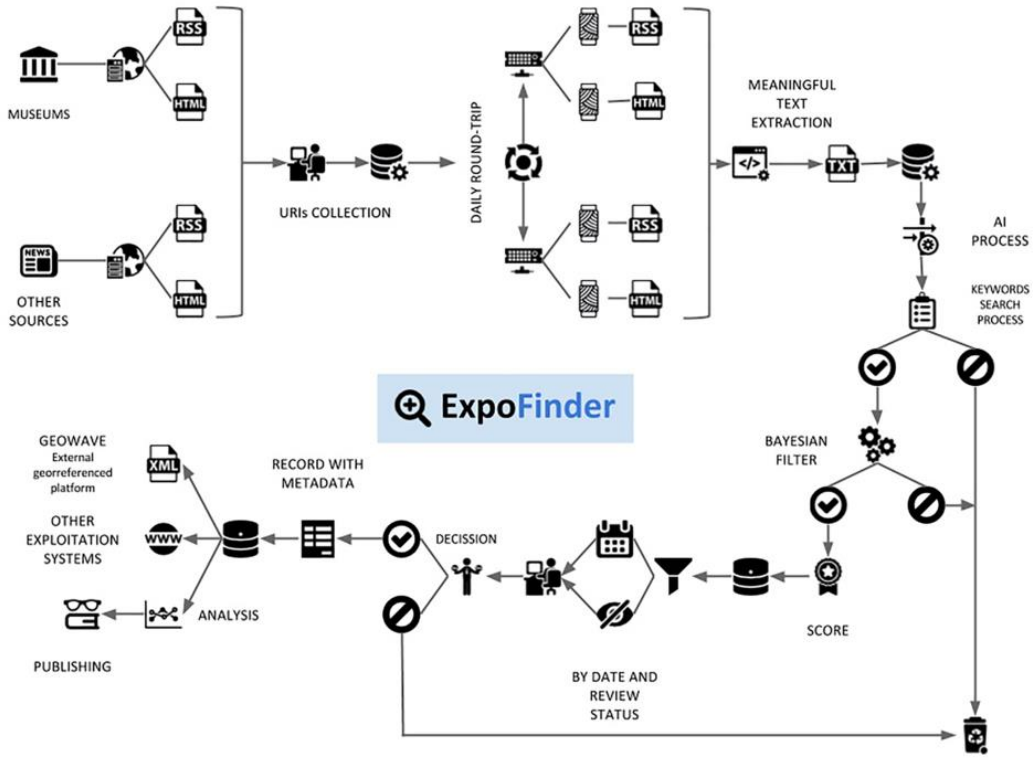

Fig. 2 Simplified Expofinder's flowchart ©Antonio Cruces Rodríguez

There are various web mining strategies that might have been used for the purposes mentioned earlier. Naturally, semantic web technologies such as the use of preexisting ontologies would have been highly efficient, but when specific domains are being dealt with, it is a more complicated issue since not all digital sources have had semantic web standards implemented or are built on a semantic representation of the knowledge that can be used in data mining processes such as we have already mentioned for the case of art exhibitions. Therefore, more viable strategies (in that they do not require the web sources to be previously structured) had to be employed such as web text mining techniques in which the use of word corpuses and specific vocabularies and terminologies play a fundamental role in extracting knowledge.

To this respect, two possible approaches were considered. This includes a more automated one, which involved defining extraction patterns to transform the extracted information into the desired output formats, and a semi-automated one, which provided the human editor with a set of texts selected on the basis of their closeness to indicators of relevance to ensure an easier filtering process. Whereas academic projects such as XWRAP Elite[16] and Wargo (Raposo *et al.*, 2002)[17] have adopted the latter solution, other projects have been based on automated learning techniques such as DEiXTo (Kokkoras *et al.*, 2013), which is based on induction and automatic data extraction. However, given the peculiarities in terms of expression and the flexibility of the idiomatic structures used to describe art exhibitions on the web, we finally decided to choose a mixed platform that combines the potential of indicator tracking systems

for text, the heuristic mechanisms characteristic of some approaches (such as Bayesian classification) and human supervision by well-trained personnel with a suitable academic background.

To preselect potential relevant texts related to art exhibitions, the Beagle + Expofinder system applies two statistical filters.

(1) A resource based on the intersection of a set of 'positive' and 'negative' keywords to which are assigned a proportional weight based on the shortest path determined by the Bellman–Ford algorithm[18]. This filtering process involves the assumption that the shortest path between significant elements of a text determines its relevance in regard to a particular topic, which, in our case, are 'art exhibitions'. In other words, Beagle detects coherent lexical chains, which indicate where the central topic of a specific text resides and seeks the shortest path among them. These lexical chains are similar to the concept of *hreb* defined by Tatar *et al.* (2013)[19]. Consequently, the oper-ational system of Beagle is based on the detection of word-hrebs and/or phrase-hrebs from a predefined list of lexemes weighted according to numeric values (Table 2). This list of key lexemes, classified by language[20], was built after analyzing the statistically relevant presence of certain words in those texts that inform about or describe art exhibitions.

(2) The other mechanism is a heuristic filter that uses a Naive Bayes classifier to calculate the probability that the text captured by Beagle is actually valid regarding art exhibition information[21]. The specific process is as follows (Fig. 3): This list of keywords is matched with the content—through the XML or HTML structure—of the text to detect hrebs. Based on the Luhn assumption and the IDF by Spärck Jones, the TF–IDF (Term Frequency–Inverse Document Frequency) statistic is applied using the binary weight for the TF and the unary weight for the IDF. Using all of the located hrebs with their values, a graph is built, where the hrebs are the nodes and the value of each node represents the value of its incoming edge (see Fig. 4). Assuming that the shorter route based on minimum weights of each possible route will represent a greater affinity of the text with our topic (art exhibition), a shortest path algorithm (in our case, Bellman–Ford) is applied. The final value of the text will be the sum of the weight of the shortest paths. The result is compared with the list of valid lexemes by the use of intersection of arrays. If the total value of the intersected arrays equals or exceeds the minimum preestablished threshold, the text is considered potentially valid.

Table 2 Example of the predefined English list of lexemes used by Beagle in the preselection process of texts with indication of their assigned weights

Base words	Weight
Curator	2
Curated	2
Curate	2
Gallery	1
Gallery owner	1
Artist	2
Artistic	2
Exhibit	1
Temporary	2
Installation	1
Installations	1
Assembly	1
Assemblies	1
Dissemination	1
Disseminations	1
Sculpture	2
Architecture	1
Painting	2
Carpentry	1
Precious metal work	2
Ceramics	2
Author	1
Exhibition	2
Exhibitions	2
Exhibited	2
Communication	1
Communications	1
Show	1
Shows	1
Competitions	1
Competition	1
Museum	1
Museums	1
Fair	1
Fairs	1

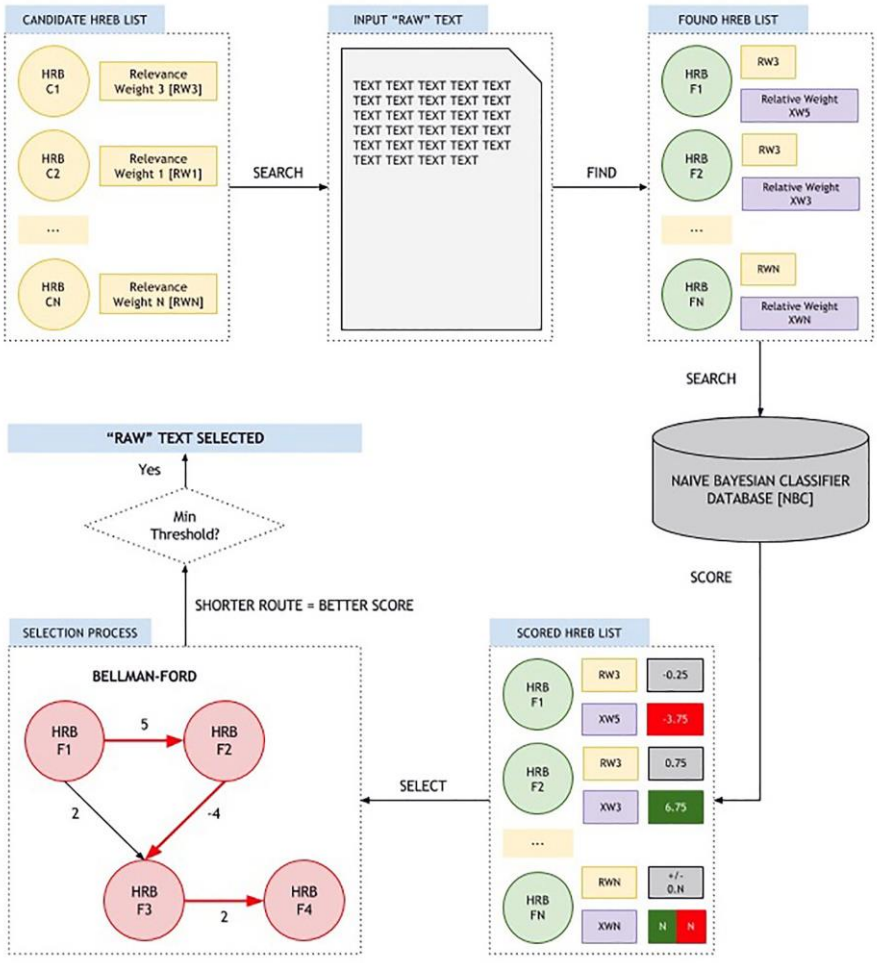

Fig. 3 Simplified run cycle operated by Beagle for preselection process of texts ©Antonio Cruces Rodríguez

It is important to notice that the Bayes classifier is not a pure discriminatory mechanism, but the source that guides the human editor in the task of deciding whether or not the information captured by Beagle is valid. The human editor is, then, who takes the final decision of integrating or not the information in the Expofinder system (see Fig. 2). At the same time, this process is the learning mechanism used by the system to progressively refine its ability to capture information about art exhibitions. Each rejection or acceptance by the editor refines its perceptiveness.

To sum up, the entire process is configured by a cyclic algorithm. Beagle captures content from Web sources, while ExpoFinder, through *ad hoc* designed algorithms, computes the percentage of the weight that the keywords have within the total text to discriminate whether such text contains relevant information about art exhibitions. It then makes a 'proposal' that the human expert must validate or reject depending on their relevance to the topic. This process, in turn, is the learning mechanism used by the system to progressively refine its ability to capture information about the art exhibitions.

At this point, some clarification about the right issues must be made. To this respect, it is important to understand the operating logic of Expofinder since it is not a harvesting system using crawling methods but an analysis tool that searches relevant information about art exhibitions in two types of

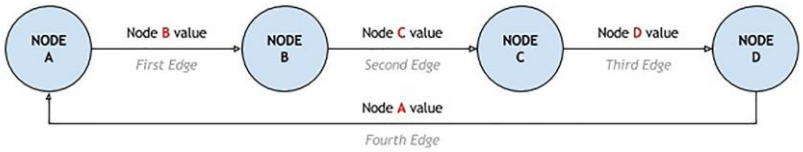

Fig. 4 Calculation of routes operated by the Naive Bayes classifier ©Antonio Cruces Rodríguez

sources (HTML and RSS/Atom) and puts it into value through specialized human intervention. The Beagle device detects web sources that can potentially convey information about art exhibitions. The information contained in these sources undergoes a datification process by human editors who are responsible for structuring, completing, and semantically enriching the primary information according to a specific data model to which we will return in the next section. Therefore, the data set that Expofinder makes available to users is never the data captured automatically from other resources. This also means that the data set provided by Expofinder does not exist as such in other sources of information or data repositories. Consequently, Expofinder uses the web sources for what they are: sources from which to obtain information that is subsequently subjected to a process of structuring and semantic enrichment. Yet, it must be also specified that in the Exhibitium Project, we work with sources whose contents are not protected by legal rights or sources that can be used with the corresponding citation. In addition, all the original sources of information that have served as the basis to build the data records are always conveniently credited.

2.1.3 *Transforming information into data: Domain model and data structure*

As mentioned earlier, Expofinder is also the system that stores in a structured format the heterogeneous and commonly semi-structured or unstructured information conveyed by the web sources found by Beagle. Therefore, once human editors have discriminated the texts that contain valid information about art exhibitions, the process continues with the task of structuring, normalizing, and enhancing such information in data form. Given that the information captured by Beagle is hardly comparable due to the disparity of the web sources examined, it was decided that developing a domain model and a specific data structure model would serve as a tool to transform such information into homogeneous data to produce a consistent data set.

Three dimensions were embraced, including the definition of a domain model with specification of types of classes and relations, the definition of a data structure model for each class, and the definition of controlled vocabularies to enrich the records with specific metadata. However, the search of previous references on which to sustain this task revealed what we have already signaled, which is the great diversity and heterogeneity that define the processes of description and documentation of art exhibitions and the difficulty of finding standard description models with enough granularity.

Regarding the domain model, our first condition was that it had to respond to the network theories, which is the theoretical framework that governs the Exhibitium Project. The Exhibitium Project approaches art exhibitions not as discrete entities that can be analyzed individually but rather as complex systems that emerge from the relations established between multiple and heterogeneous actors, which configure different and changing networks among themselves (Latour, 2005; White, 2008; DiMaggio, 2011; Schich *et al.*, 2012; Barabási, 2016). The concept, then, of art exhibition that underlies the Exhibitium Project is modeled by its nature as a social device, as a channel that makes viable relations, in other words, as a key piece in the configuration of the artistic field as a complex system. That is why we were clear from the beginning that our formalization of the domain should be represented in a graph following an ontological model. Therefore, although in the Expofinder system the exhibitions are described as entities in

themselves, the accent has really been placed on the set of associations that the actors involved in the art exhibition domain establish with each other.

Having said that, it was practically impossible to find a previous ontology specifically developed to model the field of art exhibitions. It is true that the Conceptual Reference Model (CRM) produced by International Committee for Documentation of the International Council Of Museums (ICOM–CIDOC) (Le Boeuf *et al.*, 2017) offers a formalization at a high level of abstraction that can be taken as a basis, so that the exhibitions could be under- stood as subspecies of E2 (Temporal Entity) and, more specifically, as a subspecies of E7 (Activity), but this requires a greater specification, which is not usually the case in those projects where the ICOM– CIDOC model has been used as the main codifying system. For example, if we observe the implementation carried out by the British Museum, one of the institutions that most intensively have applied the CRM of ICOM–CIDOC (2012) for codifying their collections, we notice that the information about exhibitions is not expressed in terms of CRM and is not coded. The information about art exhibitions is expressed as types of 'has note', which the CRM serves precisely to add information that does not require to be structured.

It is not the case, though, of the Museo del Prado, which has developed a representation for art exhibitions as a subgraph of its Digital Semantic Model (MSDP) (see Fig. 5). This representation, although specifically conceived to respond to the informational needs of the Museo del Prado's website, is the one that most closely fits the requirements of the Exhibitium Project. Nevertheless, when we started developing our model during the first semester of 2015, the MSDP was not available, so we could not use it as a reference.

We also noticed that the standard description models developed for recording and documenting cultural collections had paid partial attention to the exhibitions as cultural phenomenon with its own specificity since traditional systems, such as the Categories for the Description of Works of Arts (CDWA), its derivative Cataloging Cultural Objects (CCO), and the Dublin Core (DC), are mainly oriented to objects and artifacts. Despite this, in recent years, both Schema.org and Lightweight Information Describing Objects (LIDO) of ICOM (Coburn *et al.*, 2010) as well as the Europeana Data Model (EDM) do contemplate explicit structures for 'events' that might be applied for the description of exhibitions through specifications and extensions.

Therefore, taking into account this scenario, we decided to develop an *ad hoc* domain model and data structure that would respond to the requirements of our project. Complementary, we built a data mapping specification to establish the differences and equivalences between our own model and the already existing standards.

Figure 6 shows a simplified representation of the domain model that supports the Exhibitium Project formalized in a graph. The structure is articulated around the six basic elements that constitute the core set of concepts that define the art exhibition domain, including entities or institutions, people (individuals or groups), exhibitions, inscription de- vices (outputs of all nature that document and/or expand the exhibitions), companies involved in the production of art exhibitions, and featured art- works. As it can be observed, the emphasis has been placed on the diversification of the types of relations established between the classes as well as greater granularity of their properties, which thus provides enriched semantics to the data set (see also Fig. 7). The correspondences with the CRM–CIDOC's classes are indicated in the model, which will serve as the basis for its formalization in an OWL ontology on which we have recently begun to work. The most important conceptual difference concerns the notion of 'actor'. Therefore, although the CRM–CIDOC distinguishes a specific class of Actor [E39] as differentiated from Thing [E70] as a subclass of Persistent Item [E77] in turn as differentiated from Temporal Entity [E2], in our model, there is no specific class of 'Actor' because it is understood that all entities, whether human or non- human, permanent or temporary, are able of functioning as 'actors' in the Latourian sense of the term (Latour, 2005) (as agencies capable of provoking actions and/or transformation processes). Meanwhile, the most significant difference

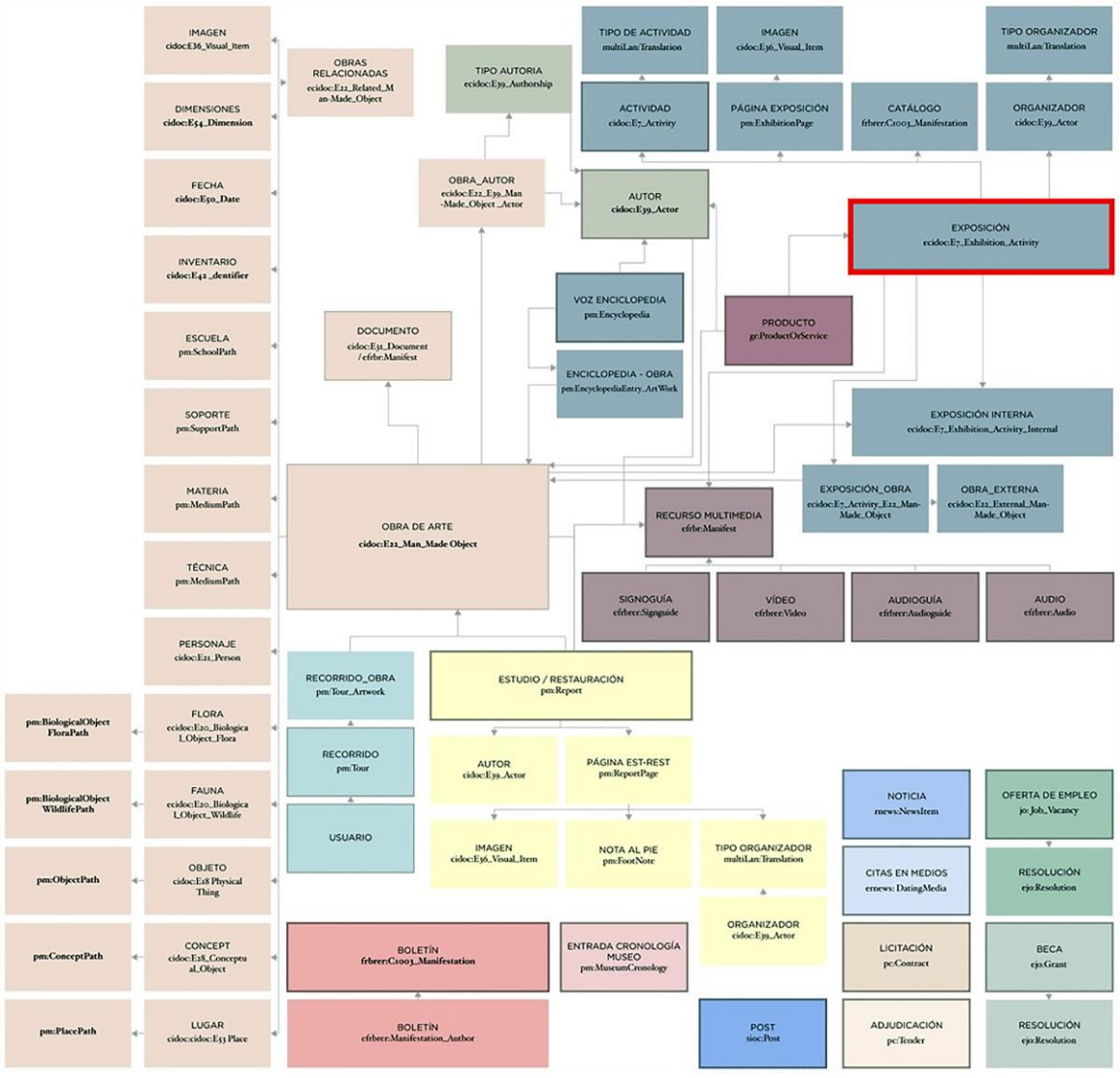

Fig. 5 Semantic Digital Model of Museo del Prado represented in a graph. Available at: https://www.museodelprado.es/modelo-semantico-digital/modelo-ontologico (accessed 30 January 2018)

with respect to the subgraph integrated in the MSDP lies in the direct relation established in our model between the exhibitions and the featured art- ists as well as in the inclusion of the curators as instances of the People class (‘actors’ in the terminology of the MSDP).

In the Expofinder system, each class also has a specific data structure although here we will only refer to the exhibition’s one, which is shown in Fig. 7. This data structure comprised three key components, including basic properties for the description–identification of the exhibition, the relational network, and properties that provide an enriched description.

Table 3 shows the metadata standards intellectual mapping, including the most common standards. In this case, the difficulties that we found were fundamentally due to issues of granularity

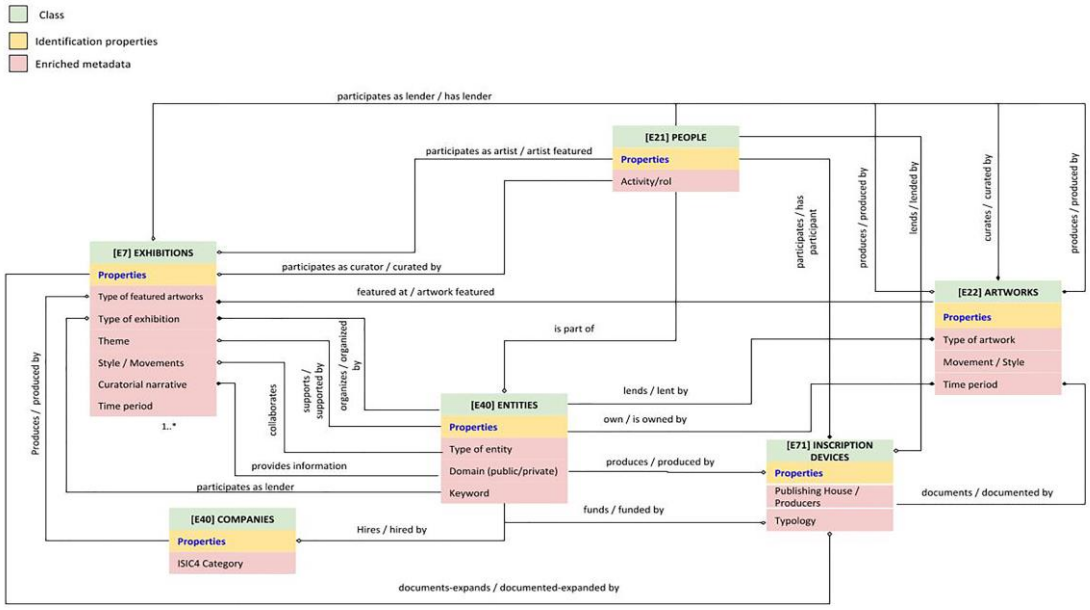

Fig. 6 Simplified representation of the Expofinder's domain model formalized in a graph (© Exhibitium Project / Nuria Rodríguez-Ortega, Antonio Cruces Rodríguez)

and specification, given that the categories codified by these standards are too generic to describe the exhibition domain in its entire complexity. Thus, for example, it has been necessary to specify the type of role played by the actors ('people' and 'entities' in our terminology) to grasp their various roles (curators, artists, lenders, organizers, supporters, museographers, etc.). In our model, this information is expressed through the specification of the relations ('organized by', 'curated by', etc.). Other problems have been of conceptual order to the extent that it was not possible to find exact equivalences. For example, should be the artists featured in an exhibition considered as creators, contributors, or what exactly? It may make sense for curators to be considered creators, but given that an exhibition is a collective project in which museographers, designers, and others also participate, should they be considered equally co-creators? Likewise, it has been necessary to add specifications to distinguish between geographical places and institutional places (venues), to include a substructure to establish differentiated connections between itinerant exhibitions and satellite exhibitions and to incorporate properties not yet considered such as those corresponding to 'narrative or curatorial discourse'.

To add value through the use of metadata, controlled vocabularies and specific authority lists were also developed. Again, we found a considerable lack in this respect. While numerous thesauri and vocabularies exist to standardize certain aspects of the artistic domain (types of objects, styles, materials, periods, authorities, etc.), there are still some gaps that directly affect the field of exhibitions such as their own typological categorization since there are no standardized categories and the use of terms is often ambiguous such as, for instance, the use of the terms 'retrospective' and 'anthological'. The exhibition themes, the discourses, and narratives that underlie curatorial design, which must be differentiated from the type of objects or styles/movements featured in the exhibition, are also needed for a categorization process. Consequently, to develop the vocabularies needed, a two-fold procedure was implemented, which involve establishing and specification of controlled vocabularies for those aspects for which previous standards existed (see Table 3)[22] and building *ad hoc* vocabularies for those aspects for which no reference model existed previously. In this respect, the

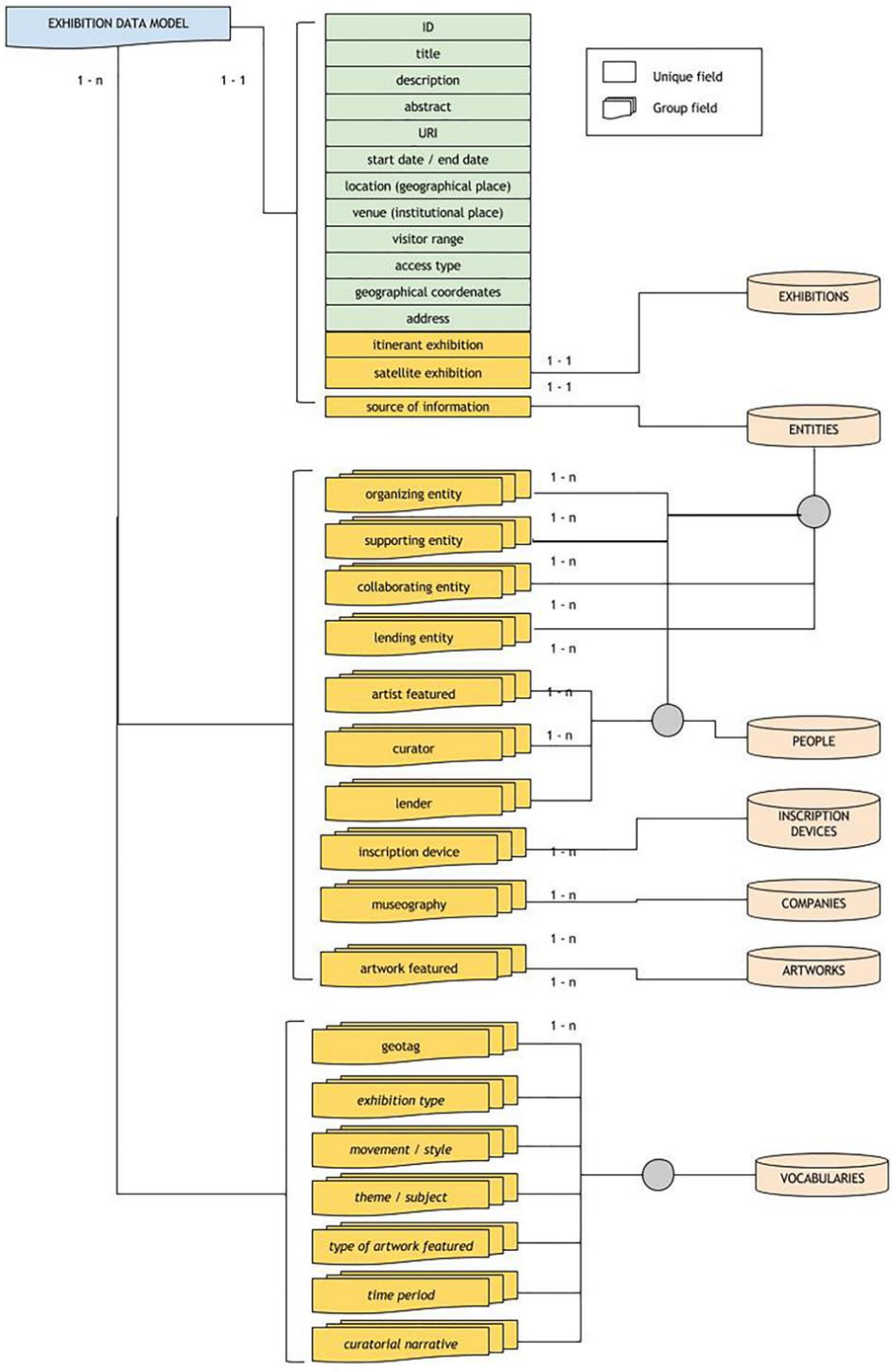

Fig. 7 Data structure for the ‘exhibition’ class (© Exhibitium Project / Nuria Rodríguez-Ortega, Antonio Cruces Rodríguez)

Table 3 Metadata standards intellectual mapping

Exhibitium	Schema.org	DC	CDWA /CCO	LIDO (events)	EDM	Control vocabularies used
ID		identifier		event ID		
URL	URL					
title	name	title	title	event name		
description	description	description	description/ descriptive note	eventDescriptionSet		
abstract		abstract				
start Date	startDate	date		earliestDate	begin	ISO 8601 date format
end Date	endDate	date		latestDate	end	ISO 8601 date format
location (geographical place)	*location*	*location*	*Place name*	*eventPlace*	*place*	Geonames
headquarter (institutional place)				*eventPlace*	*place*	VIAF
visitors range						
access type						Exhibitium
geographical coordinates						
source of information (_information provided by_)		source	source	resourceSource (Resource Elements)		
organizing entity (_organized by_)	organizer	*contributor*		*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
collaborating entity (_has collaborator_)	*actor*	*contributor*		*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
sponsor entity (_sponsored by_)	funder sponsor	*contributor*	commissioner	*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
artist featured (_features artist_)	*actor*	*contributor?*	*creator?*	*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
curator (_curated by_)	*actor*	*creator?*	*creator?*	*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
museography company/ (_produced by_)	*actor*	*creator?*	*creator?*	*actorInRole (role Actor)*	*agent (was present at)*	Exhibitium
lender (_has lender_)	*contributor*	*contributor*		*actorInRole (role Actor)*	*agent (was present at)*	ULAN/VIAF/Exhibitium
inscription device: publication, related produced resource and material (_documented/expanded by_)	*recordedIn? / review*	*IsReferencedBy?*	*source?*			
Artwork (_features artwork_)	workFeatured			thingPresent / displayObject	*isRelatedTo*	
exhibition type	*additionalType*	*type*	Work type	*eventType*		Exhibitium
movement/style			style/periods		Concept	AAT / Exhibitium

(continued)

Table 3 Continued

Exhibitium	Schema.org	DC	CDWA /CCO	LIDO (events)	EDM	Control vocabularies used
period (time coverage)		Coverage.temporal	style/periods	periodName	TimeSpan	AAT
type of object (present at exhibition)		type		objectWorkType (Object Identification Elements)	Concept	AAT
subject coverage	*about*	*subject?*	*subject?*	subjectConcept*? (Object Identification Elements)	Concept	AAT / Exhibitium
curatorial narrative and discourse					Concept	Exhibitium
geopolitical coverage		Coverage.Spatial				Geonames
associated itinerant exhibition (_has itinerant exhibition_)				*relatedEvent / relatedEventRelType*	*isRelatedTo*	
associated satellite exhibition (_has satellite exhibition_)	*subEvent / superEvent*	*isPartOf/hasPart*			*isRelatedTo*	

Note: Properties that are expressed in terms of relations are specified in parentheses. The broader properties that have needed a greater specification are marked in italic. The interrogations indicate the properties that have posed problems of conceptual equivalence.

main effort was concentrated on the building of the vocabulary for art exhibition discourses and narratives, which is still under development.

Regarding the architecture of the system, it is important to notice that the DBMS is also based on the WP structure. In the WP structure, the post is a simple content element that Expofinder uses as base to build the data records. These elements differ from each other by a table column called post type. In Expofinder, each post type corresponds to a type of record (exhibitions, institutions, people, artworks, and catalogs). It is also possible to assign some additional information (metadata) through the post-meta or custom fields to a particular type of post. This is the mechanisms that has been used to aggregate metadata to the classes of the Expofinder conceptual model. Finally, in WP, a taxonomy is the functionality that allows to group posts of a particular type. In Expofinder, the WP taxonomies have been used to build the classification system where they have been used to classify each post (or record) as belonging to a particular logical hierarchy. For example, this includes exhibition typology, type of artifacts, periods and styles, etc.

2.1.4 *Work team and task distribution*

The organizational structure of the project relies on the system workflow and is intended for optimizing the processes associated with the consolidation of the information extracted from the web sources. The structure is based on the distribution among four teams, which are outlined below:

(a) General supervision;
(b) System development and administration;
(c) Recording and verification of data; and
(d) Review of the quality and integrity of data.

The list of the actors involved in Expofinder is shown below (Table 4).

The smooth operation of the different teams depends entirely on a well-planned structure that allows some actors to perform tasks that refine those carried out by others. Thus, cohesion among members is reinforced, conceptual integrity is maintained, and useless tasks such as communication loops or complex hierarchies can be easily

Table 4 Actors involved in the Expofinder system with description of their roles

Name	Description
Administrator	User who manages the server's OS and the correct internal working of WP
Beagle	Cyclical automated process that captures information from the RSS and HTML URIs, filters that which meet a minimum level to be considered valid and record it as a draft to make it available to the Recorder verifier
Recorder verifier	Users who validate the information captured by Beagle. Rejects data that may have been mistakenly filtered as appropriate due to an error committed by Beagle. The user has the permission to modify the information for which it is responsible. The user has no permissions to amend the information verified and recorded by other users of its kind. The user has permission to modify the information captured by Beagle
Reviewer	User who verifies the correct application of the rules on consistency of information; has permissions to modify the information stored by the Recorder verifier
System	EF software system
Supervisor	User that validates the information, authorizes modifications, and evaluates the success of EF; has permissions to modify the information stored by Recorder verifier
External user	Users that access the information displayed on the front-end without rights of access to the back-end
WP	WordPress API used by EF

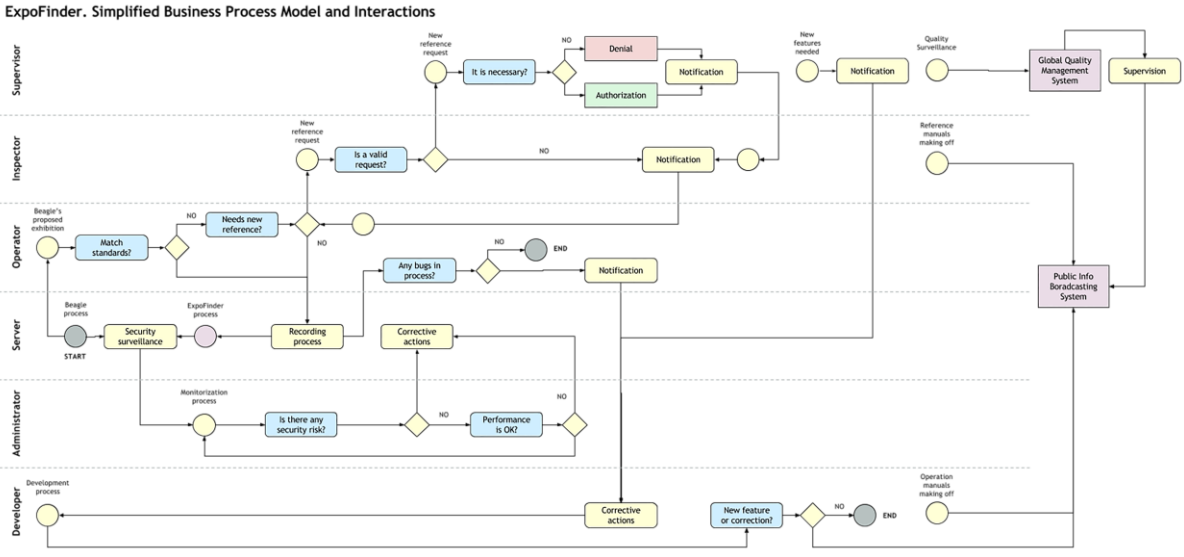

Fig. 8 Interactions among actors in the Expofinder system ©Antonio Cruces Rodríguez

eliminated. The network of categories and communications is shown in Fig. 8.

2.1.5 *Data exploitation*

One of the purposes of the Exhibitium Project is to facilitate the reusability of both the technological infrastructure and the data repository produced. For that reason, the complete Expofinder's source code is available on GitHub[23], so that it can be reused to build similar architectures or to be improved and expanded with new code developments. In addition to that, the Functional Requirements Specification document and the complete data model are published on the Exhibitium Project website under a Creative Commons license[24].

Currently, the ExpofinderDB contains more than 49.359 records, including entities (institutions and organizations), people (artists, curators, collectors, critics, etc.), inscription devices (mainly catalogs),

and companies with more than 450.000 metadata and more than 100.000 connections. This constitutes, therefore, a data repository ready to be analyzed using different strategies and techniques, which is continuously updated on a daily basis. To facilitate the reusability of data, the Expofinder's architecture also comprises a system to export data to other platforms and software where to be exploited and analyzed. This export system is governed by standard exchange rules, so that anyone interested in exploiting these data can use them for specific purposes. The only requirement is the proper citation of Expofinder as the source. Therefore, researchers and users can download the Expofinder data set entirely or previously filtered according to their interests in a variety of formats such as CSV, Excel, etc.

For example, Fig. 9 shows the graph resulting from processing the relations established among curators, artists, and institutions as they are recorded in Expofinder using Gephy[25]. In this graph, artists are rendered according to their in- degree values, the connections they receive from curators and institutions based on the exhibitions in which they have participated. Therefore, the artists with greater weight are those who are recurrently featured in exhibitions and constitute what we could call the 'classical canon' of the Spanish exhibition system. This is exactly what we discover in Fig 9 when we notice the relevant weights of Picasso, Dalı´, Miró, Goya, and Ta`pies. Simultaneously, this graph also represents the institutions and curators that contribute most to the configuration of this classic canon as it can be appreciated by observing the thickness of the edges. See, for example, the strong links between the Museo Reina Sofía and Dalí, Miró and to a lesser extent Tàpies; the significant Museo del Prado's

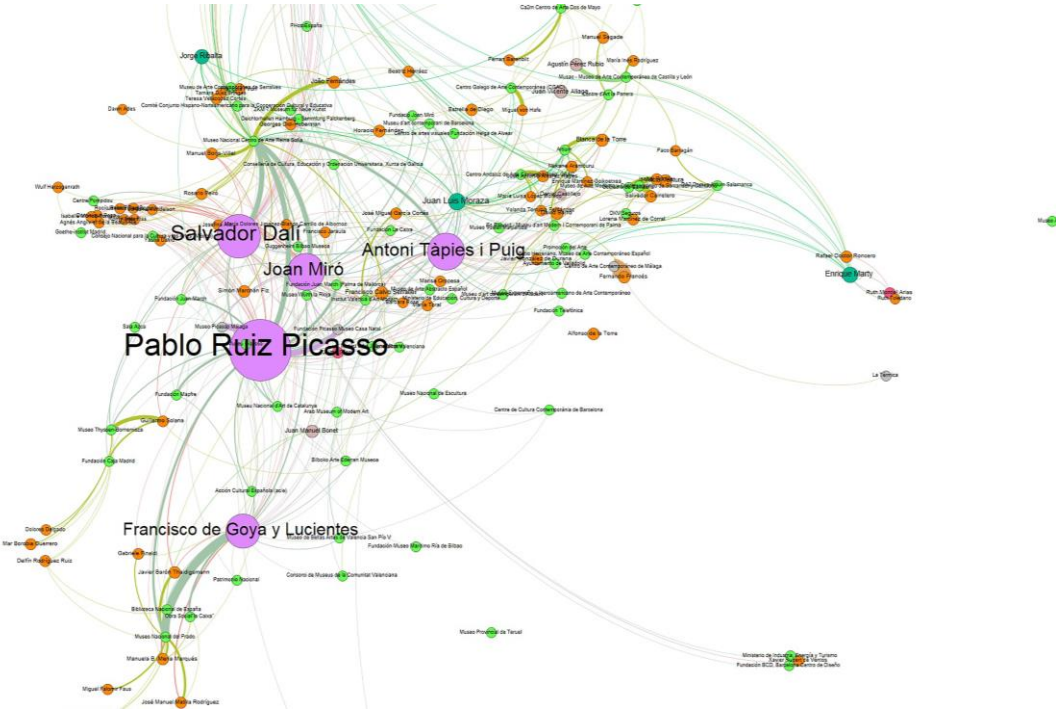

Fig. 9 Graph rendering connections between institutions, artists, and curators. Actors are represented according their in-degree values. Graph filtered by degree range 78. Scope: exhibitions held at Spanish institutions (2010–2016)
Source: Expofinder DB. Software: Gephy (© Exhibitium Project / Nuria Rodríguez-Ortega)

connection with Goya, while Picasso presents a more heterogeneous and distributed network of relations.

The Expofinder system also includes a device for the direct exploitation and visualization of data, which is available at the Expofinder's user interface. Two blocks make up user interface. The first block allows users to search for information about any record (entities, people, companies, exhibitions, etc.) and explore their metadata in detail. In this sense, Expofinder works as a traditional information system (Fig. 10).

The second block allows users to obtain data from statistics and clustering queries, which are executed by specially designed Structure Query Language (SQL) statements. Therefore, it is possible, for example, to search for artists featured in exhibitions, catalogs published, and curators participating in exhibitions (Fig. 11). This block also includes user-operated tables to create cross-referencing queries (pivotable) and to display them according to different visualizations. Just to give an example, the stacked bar chart of Fig. 12 reveals, in the first place, a curatorial practice fundamentally grounded on Spanish actors, while the international share is concentrated on just three countries: United States, France, and Germany (in this order). In the second place, it also shows the stability of this scenario since it barely changes during the period.

Although many other examples could be added, we do not want to go further on this point since the interest of this article is focused on the design and development of the technological infrastructure as well as the conceptual and data model that the Expofinder system embodies as a response to a specific need of data studies in the cultural field rather than on the eventual results that could be reached through the analyses of the Expofinder data set. Those results will be discussed in subsequent publications. What we do want to highlight is the fact that the Expofinder system completes the whole cycle, which begins with the process of detecting and capturing information about art exhibitions from web sources, continues with the transformation of distributed and unstructured information into structured and enriched data, and culminates with functionalities that allow the analysis of the produced data set to bring to light, new findings.

3. Conclusions

The motivation that underlies this article is to confront some of the problems raised in the context of cultural data studies to offer some responses. Consequently, this article is based on a fundamental premise, which is that data-driven analysis represents a paradigm of fruitful exploration due to its capacity to transform and expand cultural research toward unprecedented intellectual horizons. However, the focus of this article is not those aspects concerning the analysis strategies in themselves but the mechanisms associated with data production. As indicated at the beginning, we aspire to offer some answers to the challenge of accessing a set of heterogeneous, mass-scale, dynamic, and globally distributed sources and of transforming the content of these sources into reusable data for the creation of new knowledge and value. Once the criterion based on critical and representative selection of sources has been contested by the notion of 'massive' as a new paradigm of knowledge, the need to articulate mechanisms that allow us to obtain large data sets that can be used as primary sources for cultural research emerges as a pressing question that, in turn, entails a variety of problematic issues that must be undertaken.

First, when researching particular cultural domains, we become aware of the absence, precariousness, dispersion, and disparity that characterize data in many of them and, therefore, the limits of data-driven perspective rise as a crucial problem. The field of art exhibitions is just an example, since many others could be provided. In view of this situation, we could adopt two attitudes. First, to forget them and focus on those cultural domains with enough data sets that are to be studied through data analysis techniques. Certainly, it could be a strategic solution, but it would lead to important unbalances in the digital research on cultural phenomena. A more responsible action would be to dedicate some of our efforts to build structured data sets as an open repository to cover these

20th Century Portraits

Ficha técnica

- ID de registro: **97929**
- Fecha de creación del registro: **Hace 12 meses [domingo, 12 de febrero de 2017, 20:26]**
- Fecha de modificación del registro: **Hace 4 meses [miércoles, 20 de septiembre de 2017, 16:51]**
- Taxonomías: **Tipos de exposición.** Exposición colectiva. **Tipos de obras.** Artes visuales, Dibujos, Esculturas, Obras en papel, Pinturas. **Temas.** COMISARIO, COMPLETADAS. **Movimientos.** Arte contemporáneo. **Periodos.** Siglo XIX, Siglo XX

Metadatos

Campo	Valor
Mapa de localización	
Dirección postal	**New York City Hall, New York, NY 10007, EE. UU.**
Fecha de comienzo	**1942-12-09**
Fecha de finalización	**1943-01-24**
Localidad	**New York; Nueva York; Estados Unidos**
Sede	**Museum of Modern Art**
Rango de visitas	**< 100**
Tipo de acceso	**No disponible**
Tipo de relación	**Sede principal**
Fuente de información	Museum of Modern Art
Entidad organizadora	Museum of Modern Art
Coleccionista	**103084: Justin Thannhauser, 97797: Buchholz Gallery, 97804: Bignou Gallery, 23805: Museum of Modern Art, 102324: Frank Crowninshield, 102875: Charles B. Goodspeed, 103177: Miss Etta Cone, 97365: Paul Rosenberg and Co.**
Comisario	Monroe Wheeler
Autoría	Pablo Ruiz Picasso, Berenice Abbott, Ivan Le Lorraine Albright, Albert André, Raúl Anguiano, Milton Avery, Saul Baizerman, Balthus (Balthazar Klossowski de Rola..., Richmond Barthe, Max Beckmann, George Bellows, Christian Berard, Eugene Berman, Antonio Berni, George Biddle, Giovanni Boldini, Pierre Bonnard, Alice Boughton, Constantin Brancusi, Alexander Brook, Paul Cadmus, Massimo Campigli, Vincent Canadé, Leonora Carrington Moorhead, John Carroll, Federico Castellón, Marc Chagall, Jean Charlot, Giorgio de Chirico, Paul Cordes, Joseph Cornell, Lily Cushing, Salvador Dalí, Jo Davidson, Paul Delvaux, André Derain, Charles Despiau, Otto Dix, Louis Dlugosz, Kees van Dongen, Marcel Duchamp, Raoul Dufy, Thomas Eakins, Eliot Elisofon, James Ensor, Jacob Epstein, Max Ernst, Walker Evans, Alfeo Faggi, Donald Forbes, Jared French, Eugenie Gershoy, William Glackens, Arshile Gorky, William Gropper, George Grosz, Jesús Guerrero Galván, Hermann Haller, Minna R. Harkavy, Marsden Hartley, Herbert Haseltine, Robert Henri, Luis Herrera Guevara, Stefan Hirsch, Morris Hirshfield, Ferdinand Hodler, Charles Hopkinson, Lotte Jacobi, Augustus John, Frida Kahlo, John Kane, Edward McKnight Kauffer, Eric Henri Kennington, Ernst Kirchner, Guitou Knoop, Oskar Kokoschka, Käthe Kollwitz, Walt Kuhn, Yasuo Kuniyoshi, Gaston Lachaise, Hermann Landshoff, Marie Laurencin, Fernand Léger, Leonid Sókov, Julian E. Levi, Wyndham Lewis, L. Jean Liberte, Jacques Lipchitz, George Luks, André Lurcat, George Platt Lynes, Loren MacIver, Aristide Maillol, Man Ray, Joseph de Martini, André Masson, Henri Matisse, Alfred H. Maurer, Knud Merrild, Guillermo Meza, Kenneth Hayes Miller, Joan Miró, Amedeo Clemente Modigliani, Paul Mommer, Roberto Montenegro, Joep Nicolas, Isamu Noguchi, Juan O'Gorman, Chana Orloff, José Clemente Orozco, William Orpen, Jules Pascin, Guy Pène du Bois, Pablo Ruiz Picasso, Henry Varnum Poor, Cândido Portinari, Odilon Redon, Pierre Renoir, Diego Rivera, Auguste Rodin, William Rothenstein, Georges Rouault, Henri Rousseau, Reuven Rubin, John Singer Sargent, Concetta Scaravaglione, Elsa Schmid, Henry Schnakenberg, Kurt Seligmann, Ben Shahn, Charles Sheeler, David Alfaro Siqueiros, John Sloan, Juan Soriano, Chaim Soutine, Moses Soyer, Raphael Soyer, Eugene Speicher, Eugene Spiro, William Steig, Joseph Stella, Florine Stettheimer, Alfred Stieglitz, Rufino Tamayo, Pavel Tchelitchew, Sacha, Jacques Villon, Édouard Vuillard, Franklin C. Watkins, Max Weber, J. Alden Weir, Anita Weschler, James McNeill Whistler, Gertrude Vanderbilt Whitney, Grant Wood, Mahonri Young, Ossip Zadkine, Marguerite Zorach, William Zorach, Anders Zorn

Fig. 10 Exhibition record as displayed on the Expofinder's user interface (screenshot)

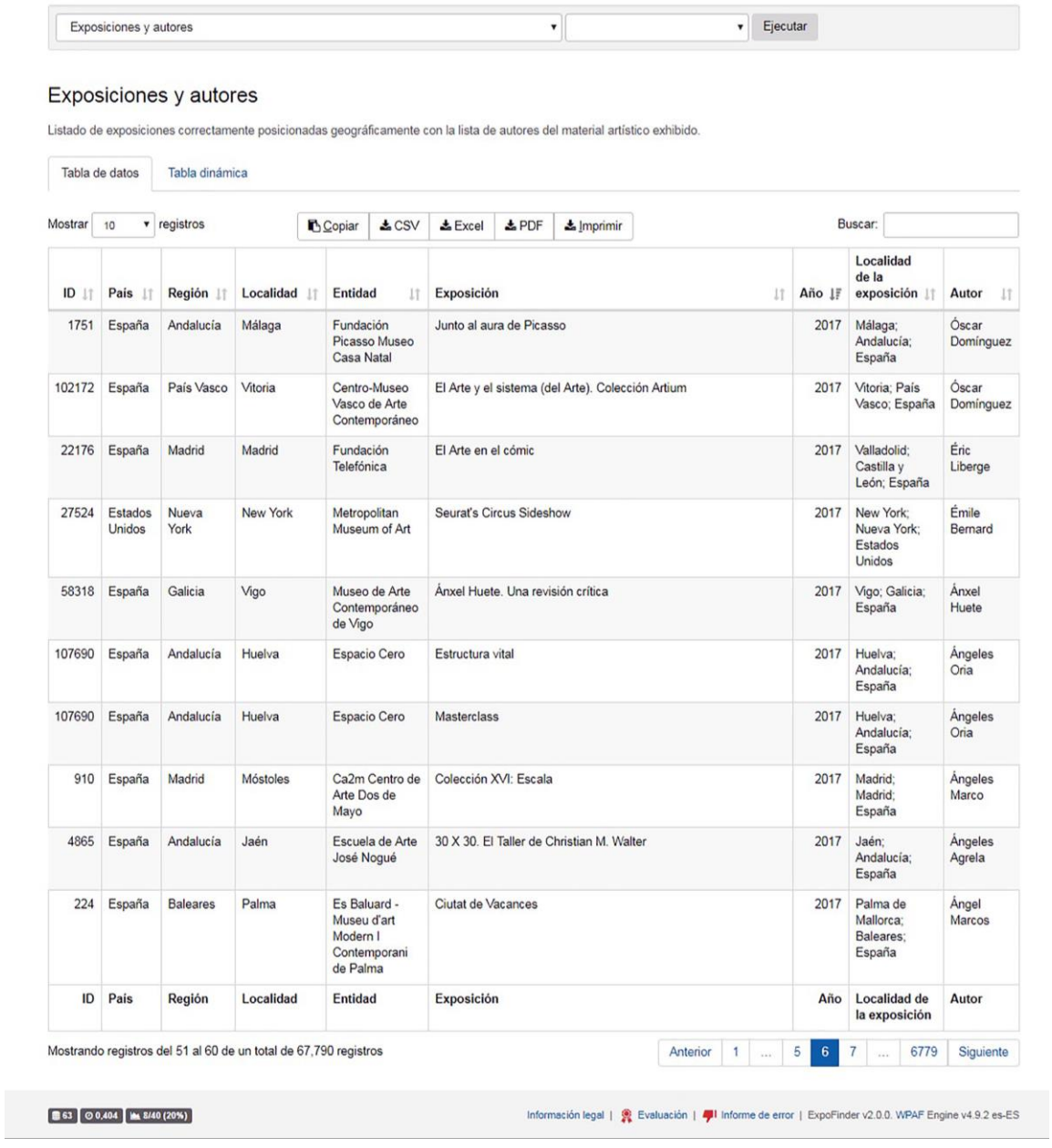

Fig. 11 'Exhibition-artist' query displayed in a table on the Expofinder's user interface (screenshot)

weak domains. We are strongly convinced that our responsibility as art historians engaged with the data-driven paradigm shift not only applies specific analysis techniques, but to detect the uncovered cultural domains and to contribute to produce new structured data. The aggregation of structured data and metadata represents one of the main tasks that we must undertake since they constitute the

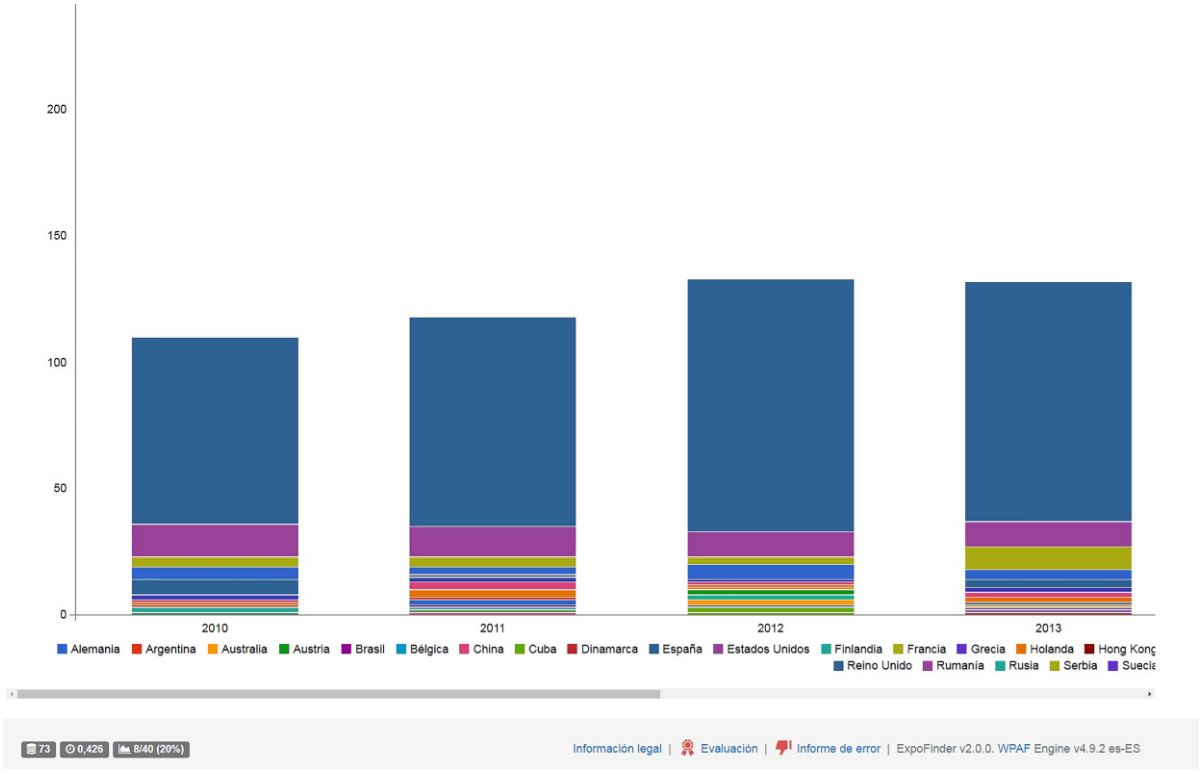

Fig. 12 Stacked bar representing annual distribution of curators performing in Spain according their nationalities (screenshot)
Source: ExpofinderDB (© iArthis_Lab)

foundations of everything else. Without them, the rest of resources, tools, analysis techniques, etc. are worthless. We need well-curated and well-structured data.

In this sense, the Exhibitium Project, using specifications and adaptations of existing ontologies and data standards, provides a domain conceptual model and a specific data structure model to codify art exhibitions until now non-existent and with vocation to become a reference for other projects focusing on this field. Given the lack of previous normalized models to describe art exhibitions as specific cultural phenomena, we consider that the one developed within the framework of the Exhibitium Project could serve as reference. In addition, one of our medium-term objectives is to transform this model into a formalized Web Ontology Language (OWL) ontology that can be used extensively by other projects and/or institutions. Likewise, in a scenario characterized by the disparity of criteria in the processes of description and documentation of art exhibitions and by the lack of consensus about which is the relevant information that needs to be preserved and codified, we consider that the very existence of a specific model could help to clarify this field. Of course, like all data models, it is an intellectual construction that involves a series of epistemological decisions and rests on a particular vision of art exhibitions as complex cultural phenomena, which can be the subject of discussion and controversy. However, we believe that the proposal of a specific model opens a fruitful line of debate as the art exhibition domain is turned into the focus of the problem.

Second, through the design and building of the Expofinder system, the question of how to deal with heterogeneous information that is dispersed, disaggregated, and unstructured has also been addressed. Naturally, other solutions could have been articulated, but we also believe that the Expofinder system, in its task of detecting dispersed and disparate information and transforming it into a structured and consistent data set, responds well to the conditions of a particular context characterized by the absence of open structured data sets and, especially, by the lack of data sets in the LOD cloud that would allow us to reuse them by computer applications in a distributed manner. Meanwhile, the ability of the Expofinder system to detect dynamic information about art exhibitions in real time guarantees that the data set is always updated.

Nevertheless, the Exhibitium Project does not advocate a scenario defined by the multiplication of similar data sets. On the contrary, it aspires to move toward structures that allow an exchange and reusability of data in a distributed space. This is the mission that we consider our future OWL ontology might fulfill to the extent that, conceived as an interoperable ontological structure, this would contribute to increase the types of LOD available, which expands the accessibility of distributed data.

We also consider that the Expofinder system, in terms of conceptualization and operational logic, represents a model susceptible to be extrapolated to other projects that require extracting, compiling, and analyzing data from dynamic, mass-scale, heterogeneous, and distributed sources and that have to confront a similar scenario. Therefore, the source code, the working methodology, the domain and data models, and the Functional Requirements Specification document are open and available for researchers and developers.

Finally, we want to emphasize the relevance of this approach focusing on the design and development of technological infrastructures rather than on the results of data analysis. From our view, it is important because this approach integrates the production of technological infrastructures and resources with all their associated issues as an essential and constitutive part of the art–historical research's concerns. In other words, approaches of this nature bring to the forefront the material conditions that underlie the production of art–historical knowledge, which align with the renewed materialist turn in the field of cultural studies. It is true that the ultimate purpose of the Exhibitium Project within the framework of which the Expofinder infrastructure that has been developed is to analyze the produced art exhibition data set to unravel the complex system of actors and strategic networks that inform this field as a key factor in contemporary cultural flows and processes. However, the findings discovered and conclusions raised from such analysis will be discussed in sub- sequent publications.

Funding

This work was supported by the Banco de Bilbao Vizcaya Foundation (FBBVA) [*Generación de conocimiento sobre exposiciones artísticas temporales para su reutilización y aprovechamiento multivalente*]. It is also part of the research project funded by the Ministry of Economy and Competitiveness of Spain [*Catálogos artísticos: gnoseología, epistemologías y redes de conocimiento. Análisis crítico y computacional*, HAR2014-51915-P] and the project funded by the Centro de Estudios Andaluces of the Junta de Andalucía [*Metodologías de datos aplicadas al análisis de las exposiciones artísticas para el desarrollo de la economía creativa*, PRY128-17].

References

Barabási, A.-L. (2016) *Network Science*. Cambridge: Cambridge University Press. [TQ1][TQ2]

Baumgartner, R. *et al*. (2009). Web data extraction system. In *Encyclopedia of Database Systems*, vol. 1. New York, NY: Springer.

Coburn, E. *et al*. (2010). LIDO-Lightweight Information Describing Objects. Version 1.0. ICOM-CIDOC.

Cruces Rodríguez, A. *et al*. (2016) WordPress as Framework for Automated Data Capture, Filtering and Structuring Processes. In *Digital Humanities 2016: Conference Abstracts*. Krako´w: Jagiellonian University & Pedagogical University, pp. 469–71.

http://dh2016.adho.org/abstracts/19 (accessed 15 December 2016).

DiMaggio, P. (2011). Cultural networks. In Scott, J. and Carrington, P. J. (eds.), *The SAGE Handbook of Social Network Analysis*. Los Angeles, CA: Sage, pp. 286–300.

Fernández, D. (2014). De la exposición de papel a la exposición virtual. Auge y decadencia de los catálogoss de arte y otros materiales de difusión de las exposiciones entre los siglos XX y XXI. In Bravo, M. C. and García, W. R. (eds.), *Las redes hispanas del arte desde 1900*. Madrid: Consejo Superior de Investigaciones Cientı´ficas, pp. 197–209.

Kokkoras, F. *et al*. (2013). DEiXTo: A Web Data Extraction Suite. In *Proceedings of 6th Balkan Conference in Informatics (BCI 2013)*, Nueva York, NY, ACM.

Latour, B. (2005). *Reensambling the Social. An Introduction to the Actor-Network Theory*. Oxford: Oxford University Press.

Le Boeuf, P. *et al*. (2017). Definition of the CIDOC Conceptual Reference Model. Version 6.2.2 ICOM/ CIDOC.

Manovich, L. (2015). The science of culture? *Social Computing, Digital Humanities and Cultural Analytics*. http://manovich.net/content/04-projects/086-cultural-analytics-social-computing/cultural_analytics_article_ final.pdf (accessed 30 March 2017).

Marías, F. (2008). Metodologías o géneros literarios y estrategias narrativas. In *XVII Congreso Nacional de Historia del Arte, Arte i memoria (pre-actas)*. Barcelona, pp. 178–9.

Moretti, F. (2013). *Distant Reading*. New York, NY: VersoBooks.

Piscitelli, A. (2015). *Humanidades digitales y Nuevo normal educativo, TELOS (Cuadernos de Comunicación e Innovación)*, junio-septiembre: 1-10.

Raposo, J. *et al*. (2002). The Wargo System: Semi-Automatic Wrapper Generation in Presence of Complex Data Access Modes. In *Database and Expert Systems Applications, Proceedings. 13th International Workshop*. IEEE.

Rodríguez-Ortega, N. (2015). La teoría y la literatura artística ante las condiciones de nuestro tiempo. In Rodríguez-Ortega, N. and Guzmán, M. T. (eds.), *Teoría y literatura artística en España. Revisiones historiográficas y estudios contemporáneos*. Madrid: Real Academia de Bellas Artes de San Fernando, pp. 1033–79.

Schich, M. *et al*. (2012). *Arts, Humanities, and Complex Networks. Leonardo ebook Series.* Cambridge: MIT.

Tatar, D. *et al*. (2013). Hrebs and Cohesion Chains as Similar Tools for Semantic Text Properties Research. In *Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2013*. Cluj-Napoca, Romania, July 4–6, pp. 1–12.

White, H. C. (2008). *Identity & Control. How Social Formations Emerge*. Princenton, NJ: Princenton University Press.

Notes

1 Consider, for example, the wide range of user-generated content related to cultural phenomena currently available on Internet.

2 This is not intended to claim that a methodological practice should be revoked. Rather, we advocate the need to think (together with traditional methodologies) about other alternatives that allow the exploration of new epistemic paradigms of our digital contemporaneity.

3 This project is being coordinated by iArtHis_Lab research group at the University of Málaga. Partners of this project are: CulturePlex Laboratory (University of Western Ontario, Canada), Techne (University of Granada), and Khaos (University of Málaga) research groups. See http://www.exhibitium.com (accessed 30 January 2018).

4 https://github.com/tategallery/collection (accessed 30 January 2017).

5 http://collection.britishmuseum.org (accessed 30 January 2017).

6 https://old.datahub.io/dataset/rijksmuseum (accessed 30 January 2017).

7 https://americanart.si.edu/collections/search/lod/ about/ (accessed 30 January 2017).

8 https://github.com/metmuseum/openaccess (accessed 30 January 2017).

9 https://github.com/cooperhewitt/collection (accessed 30 January 2017).

10 https://github.com/MuseumofModernArt/exhibitions (accessed 30 January 2017).

11 The interesting Artl@s project, codirected by Béatrice Joyeux-Prunel (Ecole Normale Supérieure in Paris) and Catherine Dossin (Purdue University, USA), should be mentioned. This project aims to analyze the circulation processes of artistic knowledge through the mapping of a large corpus of catalogs of exhibitions held over time (See: http://artlas.ens.fr/en/data- base-2/ (accessed 30 January 2018). Although research collaborations are being explored between the two

projects, the approach which governs the Exhibitium Project is different. Our focus is on the mechanisms of the production and distribution of information about art exhibitions on the Web and how this information can be reused as data to generate new knowledge. Therefore, the infrastructure developed within the framework of the Exhibitium Project scans the Web to find information about art exhibition produced dynamically by contemporary web sources. The approach of Artl@s project is rather historical based on printed catalogs. There are also differences in terms of data model, granularity, and type of information recorded.

12 While different in nature, the issues related to the recording and documentation of the sources of information on art exhibitions has been, and remains, a constant problem due to the fragility and ephemeral nature of their related materials with the exception of catalogs, which are not always produced (Fernández, 2014).

13 See at: http://wpmvc.org/ y https://github.com/tombenner/wp-mvc (accessed 30 January 2018).

14 See at: http://framework.themosis.com/ (accessed 30 January 2018).

15 Our use of the category 'museum' is broader than that officially established by Law 16/1985 on Spanish Historical Heritage, of 26 June, which is adopted by the Directory of Museums and Collections of the Ministry of Education, Culture, and Sports (http://directoriomuseos.mcu.es/). The definition of 'museum' used here, in addition to those that are officially recognized, also includes the institutions that call themselves museums and host exhibitions. This approach has allowed us to broaden the spectrum of recorded institutions that are linked to the organization and/or funding of art exhibitions. It should not be forgotten that the objective of this project is not to determine the 'museum' status of the institutions involved, but their positions as actors in the domain of art exhibitions as well as their role in the information dissemination processes. In addition to that, we did not establish an a priori categorization of the types of institutions/organizations on which we would focus our attention. Conversely, the objective was to cover the art exhibition domain as comprehensively as possible without considering previously established categorizations that could mediate our approach such as those based on the acknowledged relevance of certain institutions. Furthermore, note in the chart of the Fig. 1, the relevant volume represented in our system by the URIs of the city councils. The reason for this unexpected amount lies in our purpose of ensuring that the capture of information was as heterogeneous and distributed as possible by avoiding the concentration on those institutions that have a greater media capacity for disseminating information. In this sense, it must be taken into consideration that many local institutions that organize exhibitions do not have their own information channels, but usually they employ the channels provided by city councils and provincial councils. That is why the recording of the URIs of these institutions became a strategic factor to reduce potential biases.

16 XWRAP Elite is a project developed by the College of Computing from the Georgia Institute of Technology. See http://www.cc.gatech.edu/projects/disl/XWRAPElite/ (accessed 30 January 2018).

17 This has now become a business application: Denodo (http://www.denodo.com) (accessed 30 January 2018).

18 The Bellman–Ford (or Bellman–Ford-Moore) algorithm calculates the shortest paths from a single source vertex to all other vertices in a weighted di- graph. It is slower than Dijkstra's algorithm for the same problem, but more versatile since it is capable of working with graphs in which some of the weights of the corners are negative numbers. It is used in ExpoFinder for its weighting mechanism, which is useful to us, as we work with lists of lexemes for words employed as 'positive' or 'negative' markers.

19 'In our conception, hreb is a discontinuous text unit that can be presented in a set form. The set contains all entities denoting the same real entity or referring to one another in text, which concerns the same textual entity. One can distinguish morpheme-hrebs, word- hrebs, phrase-hrebs, and sentence-hrebs' (Tatar *et al.*, 2013).

20 There are parallel lists in Basque, Galician, Catalan, Spanish, French, German, Italian, and Portuguese.

21 In PHP there are many implementations of NBC solutions. Beagle uses a well-known one, called 'b8', by Tobias Leupold. As a result (for which the code library uses certain index tables in MySql, the database manager used by Expofinder), the analysis indicates a probability expressed as a percentage, a number between 0 and 1 or $P(A) \in [0,1]$—but with sign, being positive if the element analyzed belongs to a class or negative if it belongs to the opposite. Since it is always a number smaller than one, simply multiplying it by the final value of the shortest route chosen by the algorithm will be enough. Figures with negative signs will also have negative edge weights. That is why Beagle adopts the Bellman-Ford variant, since it can handle such arcs unlike the Dijkstra algorithm, which cannot work with weights less than 0. Thus, the lower

the value of the Bayesian classifier is, the greater the probability that the information will be valid.

22 It was decided to use the controlled vocabularies of the Getty Research Institute as reference models: the AAT (*Art & Architecture Thesaurus*) for types of objects, movements and styles; and the ULAN (*Union List of Artist Names*) for authorities. To complement these vocabularies, other sources were used, such as the Virtual International Authority File (VIAF) and the authority vocabulary of the National Library of Spain (*Biblioteca Nacional de Españaa*). The GeoNames database was also used for geographical and location names.

23 https://github.com/antoniocruces/expofinder (accessed 30 January 2018).

24 At the moment, only the Spanish version is provided, but an English version will be available shortly.

25 This graph is the result of processing more than 6,000 exhibitions held in Spain from 2010 to 2016, which includes more than 11,500 institutions, more than 17,500 agents (artists and curators), and more than 58,000 connections.