



HAL
open science

Analysing citizen-birthered data on minor heritage assets: models, promises, and challenges

Jean-Yves Blaise, Iwona Dudek, Gamze Saygi

► To cite this version:

Jean-Yves Blaise, Iwona Dudek, Gamze Saygi. Analysing citizen-birthered data on minor heritage assets: models, promises, and challenges. *International Journal of Data Science and Analytics*, 2019, 10.1007/s41060-019-00194-0 . halshs-02278798

HAL Id: halshs-02278798

<https://shs.hal.science/halshs-02278798>

Submitted on 4 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analysing citizen-birthered data on minor heritage assets: models, promises, and challenges

Jean-Yves Blaise · Iwona Dudek · Gamze Saygi

Abstract The citizen science paradigm and the practices related to it have for the last decade called a wide attention, beyond academics, in many application fields with as a result a significant impact on discipline-specific research processes and on information sciences as such. Indeed, in the specific context of minor heritage (tangible and intangible cultural heritage assets that are left aside from large official heritage programs), citizen-birthered contributions appear as a major opportunity in the harvesting and enrichment of data sets. With more content made available on the net by a variety of local actors we may have reached a moment when collecting and analysing spatio-historical information appears “easier”, with citizens acting as potential (and legitimate) sensors. But is it really “easier”? And if so, at what cost? Having a closer look on practical challenges behind the curtain can avoid turning the above mentioned opportunity into a lost one. This contribution discusses feedbacks from a research initiative aimed at better circumscribing the difficulties one has to foresee if wanting to harvest and visualise pieces of data on minor heritage collections, and then to derive from them spatial, temporal, and thematic knowledge. The contribution focuses on four major aspects: a feedback on the information and on the information available, a description grid for factors of imperfection to be anticipated, visual solutions we have experimented in order to support analytical tasks, and lessons learnt in terms of relations between academics and information providers.

Keywords Spatio-historical data modelling · Citizen Science · Information visualisation · Knowledge Discovery · Research Methodologies · Minor Heritage

1 Introduction

The notion of minor heritage refers to material and immaterial forms of unprotected cultural heritage, often products of rural societies. They are genuine signs of craftsmanship, culture and traditions, and important components of the life of former generations. Nevertheless, quite often minor heritage slips through the large heritage programs or documentation initiatives, and therefore there is a lack of appropriate knowledge, and an unconsciousness of its value. In that context citizen-birthered contributions appear as a major opportunity in the harvesting and enrichment of data about heritage assets (tangible or intangible, real-estate or movable). However actors potentially concerned – from active citizens to scientists or collection holders - should not overlook the difficulties created by the nature of the clues one will face in that application field.

More generally, minor heritage is concerned with memories, with history, and history deals by essence with ill-defined spatial, temporal and thematic data - for instance due to verbalization (*e.g.*, “probably in 1666”) or due to precision of the data (*e.g.*, “alongside a former riverbed”). Generally speaking, renewing the perceived value of such assets can help cultural actors (from citizens to “officials”) in their effort to preserve the assets and assess their significance. Just as in T. Pratchett’s *Discworld* gods get stronger when people believe in them, or weaker when they cease to do so, the understanding and preservation of minor heritage assets relies on the will of people to get engaged. It is therefore important, along with foreseeing potential benefits, to examine how “citizen science”

UMR CNRS/MC 3495 MAP
Marseille, France
Tel.: +33 (0) 4 91 16 43 42
Fax: +33 (0) 4 91 16 43 43
E-mail: jean-yves.blaise(iwona.dudek)(gamze.saygi)@map.cnrs.fr

initiatives can contribute in a sustainable manner to the acquisition and analysis of heritage information sets. Grounded on a contribution to DSAA (Data Science and Advanced Analytics) 2018 [1], this JDSA edition investigates promises, as well as challenges ahead for data scientists but also for heritage assets analysts in general.

1.1 The research's scope: an intersection of issues

Over time the "citizen science" concept tends to get blurred, and who exactly is meant when using the terms people, or citizens, can be a source of misunderstanding. Is a contributor to a crowdsourcing platform primarily a citizen, an enthusiast engaged in this or that field, or primarily "just a web user"?

As will be shown, the heart of the data we handle originates from citizens, who get involved – with or without relation with academics – and actually produce and publish on the net large data sets. In the context of this research, we shall call them simply "Information Providers" (IPs). Briefly said, the involvement of IPs appears as promising, yet demanding, in particular due to their heterogeneity:

- variety in terms of motivation, ranging from personal, private involvement of Heritage enthusiasts to a broad public sector actors (local communities engaged in cultural tourism for instance),
- variety in terms of analytical biases, ranging from systematic, inventory-like approaches (privileging here a thematic entry – 'oratories', 'old tools' - or there a geographic entry – 'all about my village') to subjective selections based on the perception, emotions of an IP,
- variety in terms of work processes, in relation with the familiarity of the IPs with information technologies at large, and notably with web publishing platforms.

Our research aims at better circumscribing the difficulties to anticipate when trying to harvest and visualise such pieces of information, and to derive from them pieces of Knowledge. To do so, we have pulled together three collections, that each push to the fore practical data acquisition, interpretation and visualisation challenges [2,3].

1.2 Priorities, and aspects left aside

We have chosen to focus in this paper on three aspects we consider as key drivers in such a research context: understanding the data as it stands, as it is "worded" by IPs; trying to make the best of it (*i.e.* attempts at deriving knowledge from it), and getting a better idea of who IPs are.

One of the intrinsic limitations of the work done is that we have privileged e-sources over any other kind of citizen-birther resource. In other words, our conclusions, if any, will not apply to minor heritage documentation in general, but to *the documentation of minor heritage as available for ordinary citizens on the web*. That is indeed a strong choice, which excludes *de facto* many citizen-birther efforts (*e.g.*, typically efforts of local societies printing a monthly news bulletin). From our point of view this choice is a reasonable trade-off between what we as academics would wish to focus on (online, massive, normalised datasets) and the current practices of IPs.

Because our research results in collecting a quite significant amount of pieces of information about heritage items distributed inside a territory, about passing elements of documentation and knowledge on, one could consider there is somewhat a natural application scenario ahead: contributing to the development of recommendation systems targeted for instance at cultural tourism. And indeed the data we collect could in theory quite match such expectations, with implicit relations both in terms of geoinformation (vicinities, thematic hiking, *etc.*) and in terms of historical linkages (touristic routes proposed on the basis of acquaintances in the typology, the craftsmanship, the customs and traditions, *etc.*).

Debates at DSAA 2018 have shown there is today a large community engaged in the design, implementation and testing of recommendations systems, including in the cultural area at large, and (as far as we could be concerned) including in data that has a geospatial layer. Typically, such research initiatives will bring to the fore concerns for the identification of Points-of-Interest (POIs) in relation with users behaviours [4] or concerns for the identification and classification of mobility patterns [5].

There are two basic reasons why we consider such an objective as premature, one in relation with the corpus concerned, one with the e-sources that document it:

- The corpus under scrutiny is strongly heterogeneous: from listed, monumental edifices to simplistic architectural artefacts or even ruins; from tangible points of interest, visible and likely to be recommended all year long, to intangible ones, occasionally topical (votive festivals, seasonal practices, *etc.*).
- The type and extent of pieces of information available about each heritage item are far from being consistent: recommending something (may it be a service - museum, shopping centre, restaurant, or an experience - concert, fair ...) implies there are descriptors widely available and knowledgeable across the collection of items. We are here far from being in a position to reach

that objective, given the heterogeneity of the corpus, and the heterogeneity of each item's "attractiveness" for decision makers.

Beyond, there is also a major concern that needs to be pointed at, in relation with potential users: the knowledge the cultural heritage community has of the profile and motivations of a potential "target audience" at this stage is far from being robust. With regards to that particular bottleneck our research does provide some clues, thanks to a profiling of IPs that helps bringing to the fore a significantly more well-founded understanding of who is likely to be concerned by Minor Heritage items. To make it short IPs are explorers rather than returners: [5] defines returners as "very regular in moving back and forth between their favourite places", and explorers as "inclined to explore new places while on the go". IPs tend to visit once and only once a given location but there are a number of exceptions (fairs, votive festivals, pilgrimages).

More significantly, only a minority of these explorers does consider for instance visiting and documenting an edifice as a primary goal of an exploration. Yet this minority is very influential as far as content production is concerned – a phenomenon that R. Baeze-Yates calls "the activity bias". By contrast most explorers do not contribute to the documentation of edifices per-se but as a side-effect of a more general blog-like comment about a hike, a local legend, *etc.* Finally, in a number of cases the pieces of information at hand when collecting citizen-birthed e-contributions are contradictory, or obviously cross-referenced. What kind of recommendation can then make sense, besides some kind of hazardous, unfounded and short-sighted clustering of items for communication purposes?

In other words, in the context of minor heritage collections, and unless we delude ourselves, there is still not enough grounded indications on what should be recommended (grids of descriptors that would be available for each and every item), and on who it should be recommended to. Once this is said, it is clear that there is in the "how should we try and share, or make the best of all that data" question something puzzling and inspiring – and simple usage scenarios like geocaching practices are part of the potential perspectives of this research, although they will remain unmentioned in the paper.

1.3 Content of the paper

This paper is essentially about feedback: we focus on practical challenges we have faced in terms of data modelling

and visualisation. Thereafter we introduce four main contributions.

Section 2 positions the research context with an emphasis on how citizen science practices meet minor heritage collections, and presents the strategy we have experimented in order to build mutually beneficial relations between Information Providers and the research program.

Section 3 details the case study, pitfalls and choices made to cope with imperfections in the information [6]. In that section we identify and exemplify factors of imperfection that we met in this very specific context of heterogeneous minor heritage documentation using e-sources.

In Section 4 we present some of the visual solutions we have experimented in order to support analytical tasks in such information sets, and highlight the gain of insight they foster.

Section 5 gives an overview of the evaluation strategies we have implemented up to now in order to measure the potential impact of the visualisations produced, as well as to get feedback from Information Providers. The most striking lessons learnt from the feedbacks will be shortly summed up in that section. We will emphasise societal challenges that this research has uncovered, without necessarily foreseeing them. We will in particular underline what can be seen as a paradox (not to say a hazard): more citizens engaged, more data available, more knowledge "at hand", but along with this more volatility, more chaos in the nature and structure of the data, and ultimately a substantive concern for knowledge transfer over time.

Finally, we shortly sum up in a conclusion section our main findings at this stage, and list some of the limitations and bottlenecks still ahead.

1.4 Summary

Questions that motivated this research can be summed up as follows:

- Who *does* know something about minor heritage items and can contribute to their documentation?
- What kind of online content do such actors produce?
- How can we as scientists deal with the heterogeneity of the data they produce, and demonstrate there is some insight-gaining possible in spite of that heterogeneity?
- At the end of the day, will there be something new learnt about these items?

2 Research Context

At first glance this research could be seen as another contribution to this general and powerful move towards Digital Atlases – in the legacy of archaeological or architectural atlases. Such initiatives encompass broad concerns that range from graphic semiology [7] to compatibility with standards (or more generally interoperability issues) [8,9]. In that general context of “formatting and interfacing heritage data sets” this research however builds on very specific choices, strategies, and issues that need to be made clear:

- The data we handle is not pre-formatted, and therefore falls outside of the classic “official records” approach to heritage items. We use heritage items as nodes (located in time and space) interlinking strongly heterogeneous e-sources.
- We focus on the potential added value of cross-examining though visual means such heterogeneous, and “imperfect” data sets.
- We are concerned with the way the information provided by citizens, as it stands today, can feed the analysis of heritage items. This self-imposed constraint delineates the type of findings expected: getting a better understanding of the information sets themselves, of what analytical processes they are likely to support, and not necessarily getting a better understanding of the heritage items themselves.

As a consequence this research claims no contribution to the creation or interfacing of heritage atlases –its scientific context is the impact of citizen-birthed information sets on scientific processes.

2.1 Citizens as information providers

With the emergence of widespread information communication technologies citizens are becoming today an increasingly important information source for diverse domains. Pioneering studies are done in geographical/spatial information thanks to open infrastructures like OpenStreetMap [10] and free web mapping services facilitate an implementation of the “citizens as sensors” mantra [11], with contributors generating content, and enhancing geographical information by pointing out their local particularities. In a decade time, crowdsourced geographical information took various shades of application - *e.g.*, voluntary geographical information, contributed geospatial information, user-created content [12], all in all, replacing the top-down tradition of geographical information production [13].

But this move renews challenges such as inconsistency issues, data validation and quality assessment requirements [14, 15]. In parallel typical citizen contributions for cultural heritage focus on basic tasks like annotating and tagging photos of archival material with keywords [16], transcription of historical texts [17], attributing longitude and latitude of objects by dynamically creating markers on maps [18]. But what one needs to harvest in the context of our application field is not limited to the above indications: what we need to harvest are personal memories, records of individual experiences, self-knowledge about a time gone by. In other words, the information we need to uncover is more detailed yet less precise; it is often subjective, unverifiable, hence a questioning of the above contribution modalities. And even if we base on the assumption that contributors are good-willing and trustworthy, an experience reported in [19] shows we ought to be cautious. In that experience a group of people was asked to say how they were informed of the 1986 Challenger space shuttle accident, a short while after the accident itself, and then again 20 years later. Differences in the reports talk for them-selves: recollections are significantly altered by time.

Yet uncovering personal, individual self-knowledge in citizen contributions is crucial if we want to preserve and share an understanding of minor heritage. Furthermore it does correspond to a trend that originates from citizens themselves who are today uploading self-knowledge and past experience via personal blogs. In other words the issue of how to record, share and interpret such contributions is raised. As an answer, we use the case study as an opportunity to test out alternative contribution modalities, that target individuals’ self-knowledge rather than solely focusing on collecting or decrypting massive information sets, and that can contribute to a better assessment of what imperfections are likely to be met.

More than ten years ago, M.F. Goodchild [11] coined the terms “citizens as sensors”, paving the way for the Volunteered Geographic Information (VGI) research community, while the Galaxy Zoo experiment and its extension the Zooniverse platform introduced a more generic concept referred to as people-powered research (and the latter has indeed been of use in Heritage Studies since then). Yet over time shades of that generic concept have clearly emerged [20]: crowdsourcing practices where citizens get engaged in order to support research activities initiated and conducted by academics, citizen science pro-

jects launched by citizens themselves, and finally collaborative science where citizens and academics co-define the research topic.

2.2 Steps towards a collaborative science approach.

With regards to those shades, our contribution primarily falls within the citizen science paradigm: the engagement of Information Providers, and therefore the data we base on, obviously existed prior to our research.

However we have initiated a series of workshops with IPs in order to move towards a collaborative approach by co-narrowing research questions. But in order to do so one has to be pay attention to potential misunderstandings between academics and non-academics, and in particular to clearly assess mutual benefits. Our strategy has been to try and demonstrate potential added-values on both sides.

This was done as a first step by showcasing the accomplishments of IPs in three different ways: each leaf of the platform (*i.e.* pages presenting one particular heritage item, reached after user selections) contains links to the information providers' platforms; one of the search modes is in fact a list of IPs, with for each of them a list of items they contribute to document; and a specific visualisation highlights the intensity of an IP's activity on the territory under scrutiny (Fig. 1).

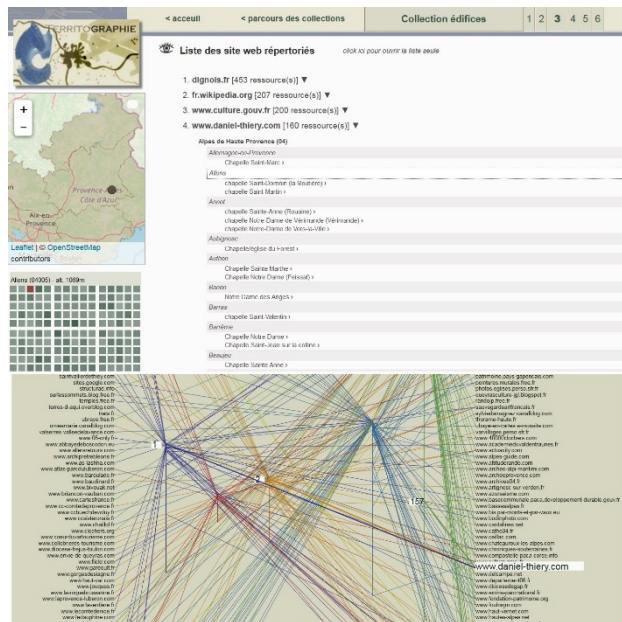


Fig. 1 Top, a browsing mode that showcases the information providers contributions – list of all edifices documented by an IP (here daniel-thiery.com, in-depth historical analyses authored by a local expert on a personal basis). Bottom, a visualisation in which the geographical distribution of edifices that a given IP contributes to document is shown. Vertical

lines correspond to the six departments in the region (intermediate administrative level). Oblique lines connect a given IP to a “number of e-sources provided” for each department. In this example daniel-thiery.com appears as quoted 157 times in one department (green), but only once or twice in two others: a clear marker of this IP's area of concern, and type of practice.

We then developed customized services for each information provider: on-the-fly production of three visualisations that show the information sets corresponding to this information provider (Fig. 2, Fig. 13, Fig. 18).

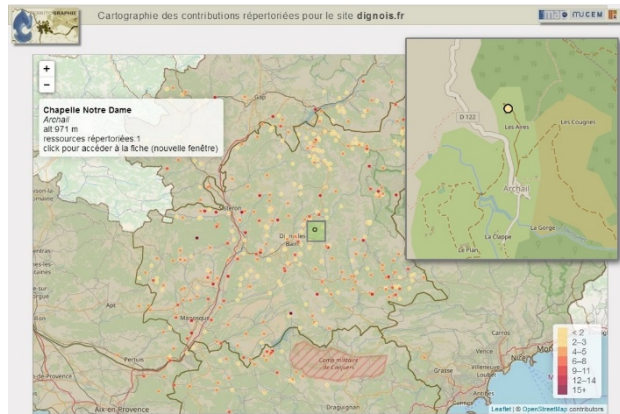


Fig. 2 A customised cartography customised, produced “on-the-fly” in order to show the pieces of information produced by one specific Information Provider (here dignois.fr). Other customised services – chronography and density/altitude analysis visualisations - are discussed in Section 4.

One should not however overlook the difficulty of aligning this strategy with the actors. Typically what we view as “services” can be totally irrelevant for this or that type of IP. Hence a necessity to try and better understand profiles of IPs, their expectations, and their work processes.

To do so we initially decided on associating each e-source we record with an indication on the status of the Information Provider – the status acting here as a preliminary, coarse grain differentiation between actors. Table 1 lists the five top-level categories of Information Providers we identified, with a definition and some examples, but also with comments on difficulties encountered when trying to classify a given IP in one category or another. The colours associated with each category will be of use in the following sections of the paper.

Table 1 The five categories of Information Providers

Public and parapublic sector

National, regional or local level public services, may they be directly in charge of heritage assets or acting as users

e.g., Public inventory services, museums, public archives, communes, tourism development agencies, etc.

A number of actors are today funded exclusively, or predominantly, by public money, but under an administrative status that is this of non-profit organisations (NPO). Considering them as public actors could be seen as going against the truth. We however do so since the work programme, and methodology, of such actors is bounded by policies of “public” decision makers.

Associations

This group encompasses a wide range of NPOs from those engaged in local history as such to those organizing events “somehow in relation”: fairs, hiking trails, cultural visits, *etc.*

e.g., Archaeological or historical societies, foundations collecting funds for heritage maintenance works, groups of citizens initiating repairs or organizing events, etc.

The ambiguity here is that in some cases an individual will act in relation with such a collective body, but with his own methodology, calendar, and sometimes objectives. There is therefore a tricky overlapping between this category and the next one.

Personal

Content produced by an individual as part of a personal commitment to promoting and documenting places (local histories, blogs), practices (crafts, transhumance), history at large, or thematic collections (rural chapels, tools, old postcards and pictures, *etc.*).

e.g., Blog of a retired Historian, collections of images harvested from personal archives, directory of places to visit as recommended by an individual, recounts of a hike, etc.

What is particular and challenging for this group of actors is their fundamental autonomy in terms of object of study, method of description, and area of concern. Furthermore, if the status of the actor is rather straightforward to establish, the info he is likely to publish is sometimes affected by an unclear lineage (see Section 3), with unsaid duplications of pieces of information extracted from other sources.

Commercial

Commercial actors can encompass a variety of profiles, with little in common besides the fact that the lifespan of the content they produce is bounded by their business practices and policies.

e.g., tourist accommodation establishments or firms selling local products who contribute as part of a communication effort, online shopping actors selling for instance images from archives, typical tools, old books, etc.

The info such actors are likely to publish is as for the previous category of actors sometimes affected by an unclear lineage. Noticeably today clear distinctions between commercial actors and the parapublic sector tends to be blurred, with for the latter an increasing demand of return on investment that is likely to impact the work practices.

Communities

All actors engaged in the publication and sharing of digital content through collaborative platforms, with users potentially acting as content providers too.

e.g., Wikipedia as such, but also wikis dedicated to a specific community like genealogists etc.

Such actors are relatively easy to identify and classify however the borderline between a commercially driven community building platform (*e.g.*, Facebook) and wikis for instance could be a topic of concern.

In a second phase, we further developed this initial grid and identified eleven profiles that correspond to subcategories (*e.g.*, community builders, bloggers, collectors, local actors, *etc.*). We then started aligning these profiles with behaviours or work practices (type of content published, thematic scope, geographical area, *etc.*). This assessment of ways of doing proved helpful in the preparation of workgroup discussions with IPs (see Section 5). It was at start a way for us to try and assure a better representativeness of the “IPs sample” (Information providers called in for the workgroup discussions). But it can also help unveiling tendencies in the material a category, or a sub-category of IPs is likely to publish, or in their respective editorial choices (Fig. 3). A formal and exhaustive analysis of the data with regards to sub-categories is part of future works we in-tend to conduct.

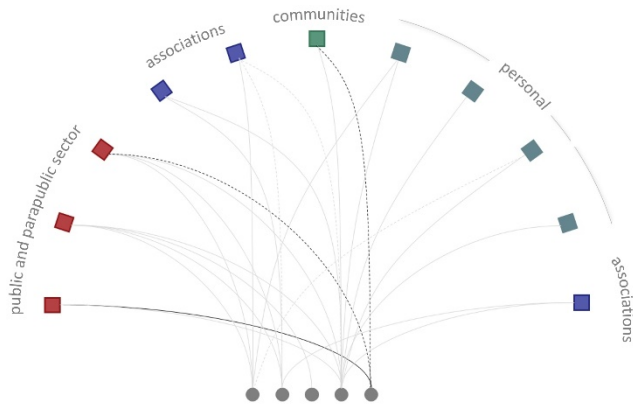


Fig. 3 Relations of IP profiles (represented as coloured squares) to publication practices. Colours correspond to “types” of actors (cf. Table 1.) The circle on the right represents “publication of raw data”. Only three profiles do consider the publication of such data sets, and two of them occasionally only (dashed line).

3 The Case Study

Our case study concerns the territory of the PACA (Provence Alpes Côte d’Azur) region, one of France’s 13 metropolitan regions, composed of 958 communes (smallest administrative layer). These communes are distributed in 6 departments, an intermediate administrative level. Communes vary in size (area and population, respectively in proportion 1:1300 and 1:3449, 2013 data).

There are most significant geographical and socio-log-ical contrasts within the region between the coastal strip, very densely populated, strongly impacted by the tourism business, a relatively barren landscape covering a large area in the heart of the region, and mountainous areas that were a century ago rather isolated, home of transhumance, and that are today turned towards tourism at large and cultural tourism in particular. The following figure (Fig. 4) gives a visual sense of how altitudes of communes, their densities of population, and their surfaces correlate.

Such contrasts are not anecdotal for a data analyst: they call his attention on patterns to look for in terms of relations (see Fig. 10) or in terms of distribution and densities (Fig. 11, 12, 13).

On the overall we pulled together three collections consisting of: 1313 rural chapels, 360 traditional farming tools from the MuCEM Ethnological museum (Musée des Civilisations de l’Europe et de la Méditerranée), and over 200 traditional crafts and professions (Fig. 5). The idea is to try and cross-examine components of the collections and to assess visually spatial, temporal, and semantic relations.

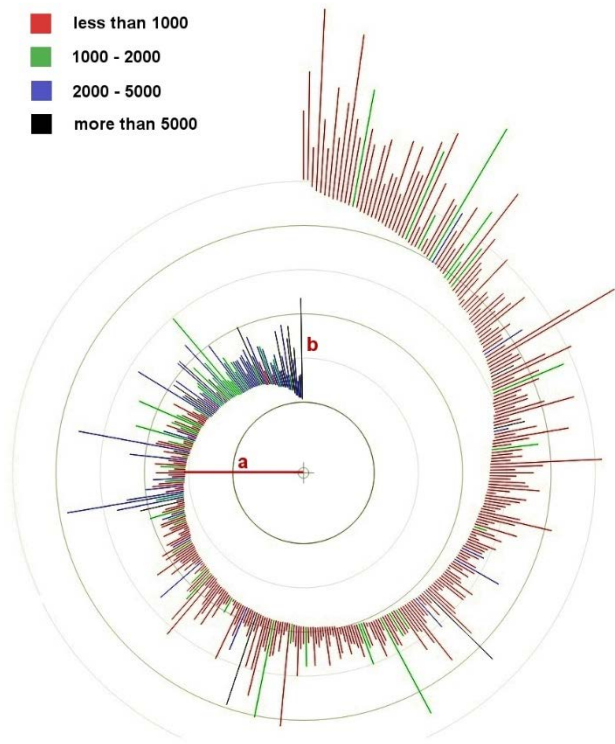


Fig. 4 Altitudes, surfaces and population of communes. Distance (a) corresponds to the mean altitude of the commune. The length (b) represents the commune’s surface. Colours indicate the number of inhabitants. The visualisation underlines a general pattern: higher communes tend to be larger and less populated. It thereby confirms the general statement that this is a region of contrasts. But this general pattern is far from being a regular one, and what is shown here indeed is that contrasts should not be seen only as a result of geographical constraints but require a fine-grain analysis, with noticeable exceptions all along the spiral.












Fig. 5 An illustration of the items in the various collections.




3.1 Harvesting the data

Although the collections are different, we extract a set of common features corresponding to the spatial dimension (classic point location, but also relations between locations as in commercial exchanges), to the temporal dimension (both linear chronology and potential cyclic behaviours), and to thematic layers underlining potential spatial and temporal concurrences across the collections.

Table 2 sums up the data harvested (or available) for each collection – relationship between among different kinds of citizen-birtherd information is far from being systematic, although an overlapping can be observed on toponymy for instance. But that overlapping that implies a critical examination in order to disambiguate homonyms, check for alternative / deprecated names, *etc.*

Table 2 Pieces of data acquired on the spatial, temporal and thematic dimensions, collection per collection

	 Spatial	 Temporal	 Thematic
edifices	 Location (long., lat.), geographical container – commune or department, altitude and orientation ^a . The notion of “spatial data” encompasses the 3D shape of the edifice – one of our fields of concern, but a topic we do not mention in the context of this paper.		
		 Time anchors: a concept developed as the hints we meet may refer to the construction of the edifice itself (“built in the 17th century”), but also to its “first mention in archives”, <i>etc.</i> A time anchor is basically the association of the edifice to a verbal or quantitative expression of a date. Other temporal indications recorded are temporal cycles of use (seasonal use for instance) and changes that occurred along the history of the edifice.	
			 Thematic layers range from architectural analysis (<i>e.g.</i> , shapes, components) to records of experiences such as pilgrimages, votive festivals, or simply travel diaries.
farming tools	 An item is, when possible, associated with a geographical container – commune or department.		
	 Two indications are available: the recording date (year of acquisition by the Museum) and when possible the period of use of the item (No indication on the cycles of use).		
	 Items are grouped by categories that act as thematic layers. Scarce indications are available on the “instructions for use”, on the making of the items, on the way they were stored and pre-served – these are the pieces of information the research primarily expects to harvest through citizen contributions.		

crafts and professions	 Association with one or several containers (communes) and when possible with a specific location (market, <i>etc.</i>).
	 Recording of periods of practice (question: “when was this craft present”).
	 Thematic layers include hints on the practices themselves, on their naming (etymology, alternative names), on tools and materials concerned, and when relevant a qualitative description of the spatial layout (mobile stand, work-shop, <i>etc.</i>).

^a Orientation of the edifice’s nave, *i.e.* of a vector running from the porch to the apse, by convention counted from the North.

The data was harvested primarily from online e-content concerning the edifices and the traditional crafts collections. For the edifices collection 3562 web pages are recorded, along with an attribute stating the status of the IP (see Table 1, and Fig. 6).

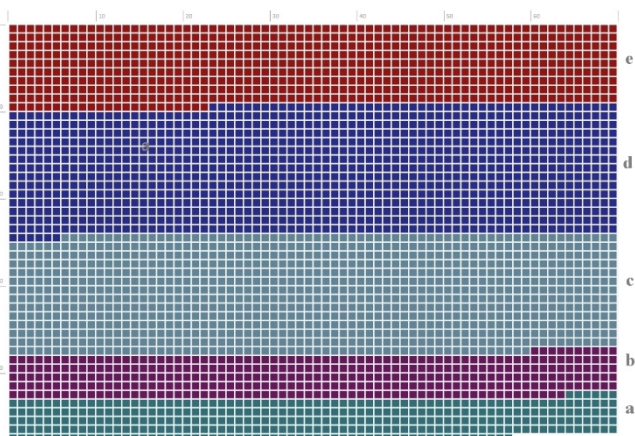


Fig. 6 Information providers’ profiles: each square in this visualisation corresponds to one of the e-sources used to document the rural chapels’ collection. The visualisation facilitates the counting of e-sources, in relation to the IP profiles: (70 rows, 48 columns, each square gives access to the corresponding URL). Colours are those defined in table 1, and are used to differentiate IP type. From bottom to top: a, internet communities; b, commercial sites; c, personal sites (individuals acting on the basis of a personal commitment to Heritage items); d, associations (NGOs) and e, public or parapublic sector (local authorities for instance).

Concerning the farming tools collection most of the data was already available in the MuCEM’s records, however e-sources were harvested in order to complement definitions or to associate items with recurrent cyclic events (fairs typically).

For the traditional crafts and professions collection a set of e-sources was also collected that document the crafts in general terms or as they developed on the territory we are interested in.

But here a traditional bibliography complements the information available, and multimedia content (still images, videos and sounds) are referenced that give a more “ethnographic” colour to the data harvested. On the overall 1139 e-sources are referenced. In both the above cases the data was harvested manually – a limitation in a sense, but a necessity in the early stages of the research in order to grab a fine-grain understanding of the nature of the data, and of the profiles of IPs.

3.2 Modelling the data: pitfalls and choices

Defining and evaluating a robust observation protocol is one of the most prominent difficulties when harvesting crowdsourced data, whatever object of study is concerned. We here base on existing data sets, with observation protocols that are hardly described, when not simply non-existent. Our focus is therefore put on the data *as it stands*, and our attempt has been to try and list the consequences, in terms of reliability, of using such data sets.

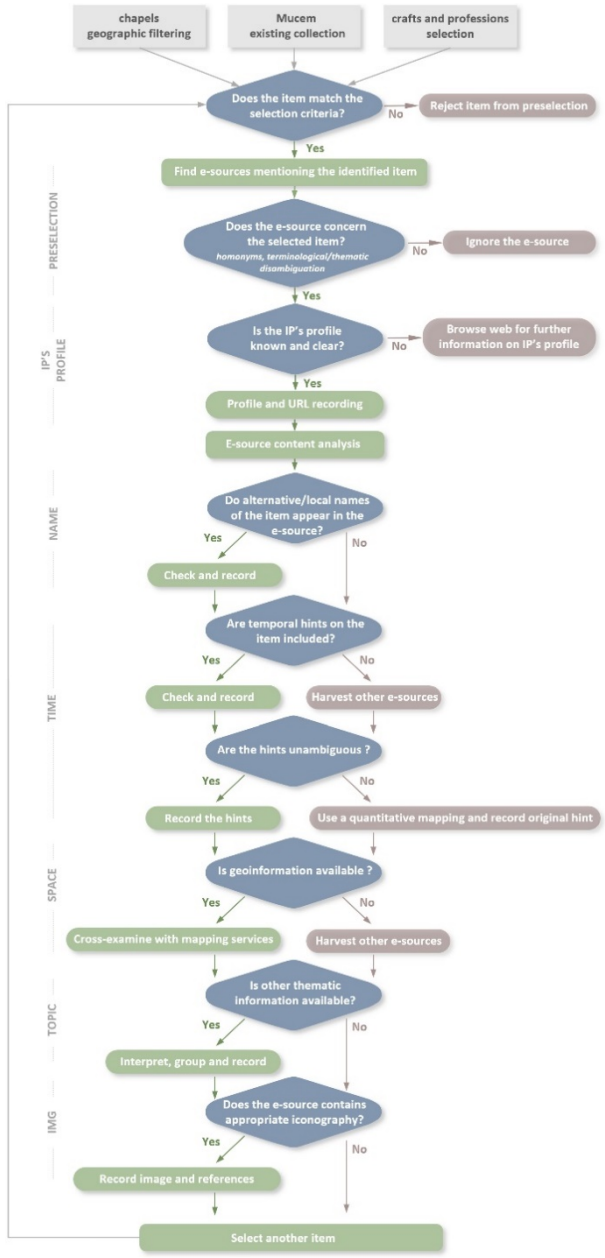
Classifying factors of imperfection in spatio-temporal or historical data has been a recurrent research topic over the past years, sometimes within the boundaries of a scientific discipline [21,22], sometimes in more interdisciplinary settings [23]. In this section we borrow notions and terms from the scientific literature and from previous research, however some of them remain today discussed. We therefore propose in table 3 a series of definitions and references in order to avoid ambiguities, along with “real-life” examples of where and how they are present in our case study. We make no claim that this list is comprehensive, yet it does pinpoint some key factors we came across, in a practical manner, and we consider it a significant part of this “feedback” of our contribution.

Most of these factors are present in the three collections, and a workable approach to “reduce” imperfections is not always within reach. It has to be said clearly that anyway in Historical Sciences reducing imperfections is in fact not a good idea at all: disagreement and open questions are full-blown components of the reasoning process. Our approach bases on the vision that what we need to convey about the data is not assertive interpretations but hints *the way they are*, yet made more readable, and shareable.

Table 3 Factors of imperfection: lexicon, references, and examples of occurrences in the data sets

<p>determinacy Whether the value of a variable [21] is known at all or not. <i>An edifice that cannot be localized (only mentioned in archives, without hints on its position), or that cannot be dated.</i></p>
<p>credibility Judgement made by the human consumer of the information about the information source. [22] <i>Association of a craft and a territory basing on the sole recollections of a witness.</i></p>
<p>approximation Attempt to come close to measuring or describing a phenomenon [22] <i>Measuring the orientation of an edifice bases on the presumption that the nave is actually straight, and the apse unambiguously positioned.</i></p>
<p>incompleteness The idea that the observed evidence is likely to only be a small portion of the whole. [24] <i>None of the collections we handle is complete, hence the necessity to be cautious in any interpretation. Further-more, as mentioned by [19], the unidentified unknowns are the worst kind of missing information, and one of the sub-goals of a citizen science approach to minor heritage can be to try and diminish the amount of unidentified unknown.</i></p>
<p>interrelatedness Source independence from other information. [24] <i>When two e-sources make a common statement, yet without quoting each other or the common initial source they based on.</i></p>
<p>currency Temporal gaps between occurrence, info collection & use. [24] <i>Temporal gaps between the period of use of a farming tool, and the moment when it was collected.</i></p>
<p>multivocality When several hints appear as contradictory. [21] <i>The literature sometimes mentions inconsistency or disagreement [24] to name such imperfections – a typical occurrence is opposite contradictory dates given for an even.</i></p>
<p>accuracy Difference between heuristic & algorithm. [25]</p>

<p><i>Transferring a qualitative indication such as “during the spring” to a given numerical interval.</i></p>
<p>imprecision Inexactness of measurement. [24]</p> <p><i>The value recorded for an edifice’s altitude depends on with what instrument / under what climatic conditions the survey was carried out: a non-systematic protocol may alter the precision of the data.</i></p>
<p>lineage Conduit through which info passed (number of steps). [24]</p> <p><i>A typical example is the recording of farming tools: the information available in the records today results from a process that involved at least the donor of the object and the museum expert, and potentially ancestors of the donor, and contemporary successive curators.</i></p>
<p>periodization Dating of a fact by reference to another one. [23]</p> <p><i>Periodization should be interpreted in a broad sense, from examples such as “in the early middle ages” or “rebuilt after the revolution” to natural facts and phenomena such as “during the crops”.</i></p>
<p>subjectivity Amount of private knowledge or heuristics utilized. [25]</p> <p><i>There is no reason to think that the way a craft was organized in one location by one craftsman as it is reported by him or his descendants is an objective testimony of what that craft is, yet we rely on such subjective hints to picture it.</i></p>
<p>likelihood At best, a stopgap verbalization in inference making. [23]</p> <p><i>Wordings such as “could have been built shortly after the Wars of Religion”; or “livestock markets probably took place in the open land close to the river bed”.</i></p>



3.3 The data harvesting and modelling workflow

Summing up in a clear-cut manner the information processing workflow is arduous: the strategy adopted can depend on the various pieces of data concerned, and on the collection being documented. The following flowchart (Fig. 7) summarises the key steps that were taken in the harvesting, filtering and recording of e-sources. Due to the heterogeneity of the items to document, and of this of the data we process, this chart should be understood as a simplification (for each type of information, and for each type of item specific steps can be needed).

Fig. 7 Harvesting, filtering and recording of e-sources: an overview of the protocol experimented.

3.4 The spatial dimension

The baseline geographical data recorded in the edifices collection are a position (longitude, latitude), an altitude, and an orientation (Fig 8). At this stage we defined a list of values (numerical scale) used to “tag” data with regards to determinacy and credibility factors (is it known at all? From what source?). The scale is used in the recording of the data, and at visualisation time in the graphic semiology. Yet this does not solve the problem extensively.

Edifices for which the location is not known at all are tagged as belonging to a container (a commune) - but what if even that indication is unavailable?

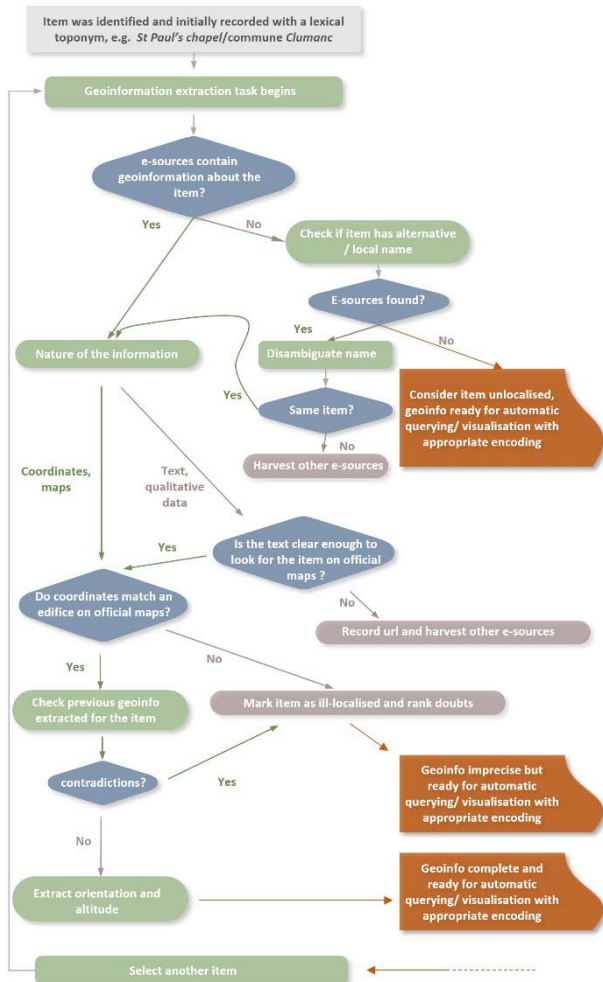


Fig. 8 A flowchart detailing, on the example of geoinformation extraction from e-sources, how information is processed, and how potential contradictions and heterogeneity are dealt with.

The data quality assessment that the process illustrated in Fig. 8 results in is original, but the process itself is, on purpose, basically manual. This is because one of the main research’s objective was for us to acquire an in-depth understanding of how Information Providers deliver information, of their work patterns and potential biases. The data quality assessment is recorded and in return impacts the rest of the processing, in particular the automatized data visualisation procedures we have implemented.

For the crafts collection we record two indications: association of a craft with one or several containers (communes) and qualitative description of the spatial layout.

But in the case of occupational travellers, a category that encompasses herdsmen moving to mountain pasture areas and travelling salesmen, we still need to propose a data model that would allow the recording of itineraries. This is not a trivial issue: recording only the start and end location is bland from the scientific point of view, adding “stop points” along itineraries (a watering place, a fair, a rest place) could make sense but leaves vast unknowns since an itinerary is hardly composed of straight lines. What is more, temporal aspects of the displacements are also very significant: a series of (Space, Time) tuples could therefore better match the reality of what we need to learn.

Finally for the tools collection defining relevant geographical data is a methodological challenge: what is the “position” of a movable object? What would be the point in recording its current position (*i.e.*, in a Museum’s reserves) from the point of view of scientific analysis? At this stage we record tuples of values: the position of the commune where an object was created and the commune where it was used. This at least paves the way for a visual analysis of exchange routes (rather basic though, see Fig. 10). However in many cases we only have an indication on the department of creation and of use, and the visualisation then is ineffective. It is plausible anyway that a more significant information would be a simple time + space assessment of presence.

3.5 The temporal dimension

As mentioned before, temporal data recorded in the edifices collection are “time anchors”. This indication is, as often in the heritage field, dramatically impacted by the multivocality and interrelatedness factors. In many cases the info we harvested is contradictory, and when not its level of independence is hard to state. However this case study also underlined a number of other potential pitfalls: determinacy (numerous edifices are simply not dated), scope (what is actually dated is not clear –presence at time T, time of construction, *etc.*), accuracy (wordings like “in the middle of the 16th century), periodization (*e.g.*, “after the middle ages”). Finally, temporal data is also strongly impacted by the likelihood factor (*e.g.*, “its construction dates back to shortly after 1516”) and by the lineage factor (for instance when dating a votive festival basing on the recollection of an individual who heard it from an ancestor).

For each edifice we record as many time anchors as read from the sources, and have defined a “conventional temporal mapping” grid that helps us transferring the verbal indication (“beginning of the 17th century”) into a quantified time slot (“1600 – 1620”) used at visualisation time. This solution was at start designed as a makeshift solution, and we are now investigating how an ontology of temporal hints verbalization modalities can help in the analysis and visualisation steps.

In both the crafts and farming tools collections what we record are time intervals during which the item is “active”, as well as hints on for instance intermittency patterns. Cross-examining the three collections showed that ultimately one of the challenging issues we need to address is harvesting and reasoning not only on a chronology, on a linear time succession of dates, but on the overlapping of recurrent temporal patterns as implicitly present in expressions such as “every summer”, “every second Sunday of July”, “during the harvesting of the grapes”. Consequently, we complement the data models that are specific to this and that collection by a data model used to collect and document cyclic (or at least non-linear) temporal occurrences (markets, fairs, votive festivals, pilgrimages, *etc.* – see Fig. 9).

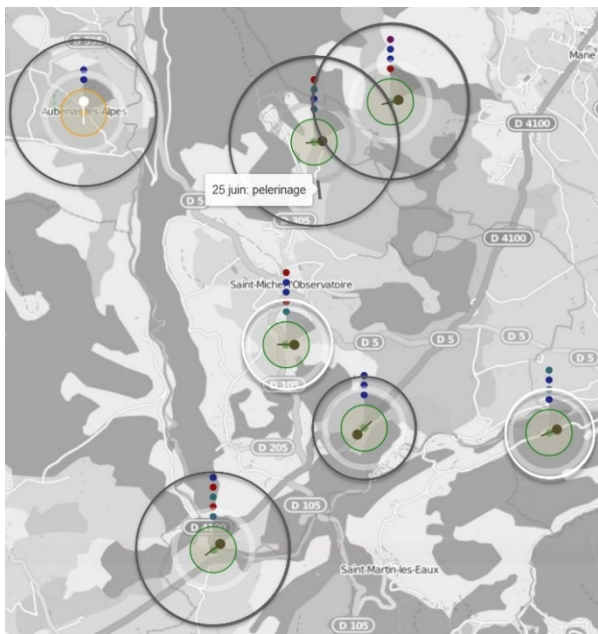


Fig. 9 A cartography on which symbols transfer information on both linear time and cyclic events. Large black or white circles correspond to linear time data: the larger the circle, the older the edifice. White circles correspond to edifices that are not dated (lacking information). One cyclic event is highlighted (a pilgrimage on June the 25th) for one of the edifices

(mouse over the little grey line opens a popup). The orientation of that little line corresponds to a moment in a year (a circle acts as a clock, representing 365 days). The symbols used also transfers the information available on the geographical orientation of the edifices and on the amount of e-sources collected for each edifice (little coloured circles aligned vertically, coloured depending on the IP’s status).

The data collected acts as a potential bridge between collections, since in a number of cases such an occurrence concerns two or three collections (*e.g.*, a fair where a craft was presented, tools were sold, during a votive festival in relation with an edifice). In parallel, that data acts as a test bench for the formalization and visualisation of cyclic temporal occurrences, including in the verbalization modalities.

4 Visualising the Data

The general objective in visualising the data is basically to cross-examine, correlate, and perform reasoning on the temporal, spatial, thematic dimensions, to spot significant patterns with keeping a concern for doubtful info. Over time we have experimented a rather large number of cartographic solutions, and of InfoVis solutions, with for each an ambition to try and answer to specific questions (a tribute to J. Bertin’s vision) such as is there a relation of the altitude to the orientation of edifices? Are there more contributors in one commune than in another one?

The inspiration behind these experimentations is at the intersection of geovisualisation, time-oriented data and InfoVis, an intersection personified by the works and legacy of C.J. Minard [26]. In this section we present four examples that we believe show the above methodological intersection is worth exploring (even with “poor” data sets in terms of quality and consistency).

4.1 A map of exchanges

Figure 10 shows a leaflet-based map [27] produced in order to explore the basic geographical data we have for the farming tools collection, *i.e.* commune of creation, commune of use. The visualisation as such, a simplification of the flow map visualisation paradigm, appears quite efficient in highlighting some tendencies, for in-stance the exchange routes between winter and summer areas concerned with transhumance related activities (SW \leftrightarrow NE). Colours (see legend), are used to represent the amount of objects created (interior circle) and used (exterior circle) in a commune.

Two communes (Arles and Névache) appear as strong “exporters” of tools. It is important however not to jump to conclusions: the data set is too inconsistent to provide a robust analysis ground – the visualisation from our point basically acts as a proof-of-concept and shows what could be gained from a larger and better documented collection.

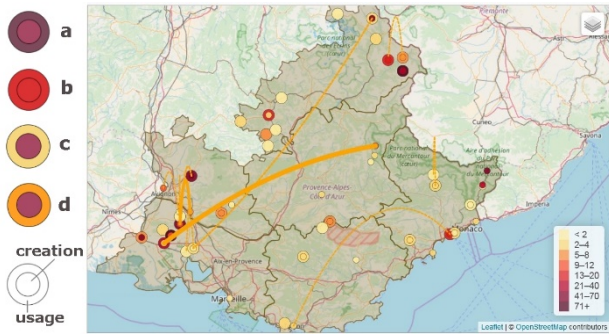


Fig. 10 A visual representation of the communes of creation and use – circles represent communes, curves are used when a tuple [creation, use] info is associated to a tool, and the thickness of the curves correspond to the amount of tools. (a - St. Véran, b - Arvieux, c - Névache, d- Arles)

4.2 Densities and altitudes

The area of concern combines very different landscapes (mountains of the Southern Alps, dry hills, a coastal strip densely populated today, etc.). What can be learnt on the spatial distribution of edifices with regards to factors such as landscape, climate, accessibility, and so forth? The following visualisation (Fig. 11, 12, 13) helps reading an interaction of parameters: area, various altitudes, density of edifices per commune / per department.

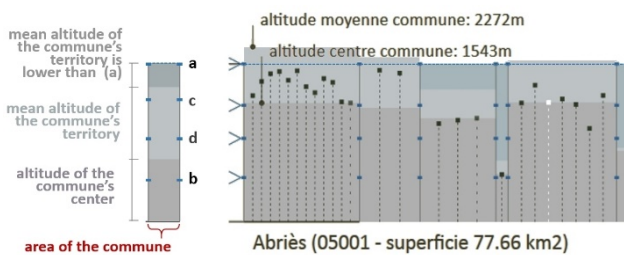


Fig. 11 Composition of the graphics: communes of a department are represented by “rectangles” aligned in a long strip. The width of a rectangle represents the area of the commune. Blue bars on the y axis represent min, max and average altitudes concerning a department: **a**) highest centre of a commune inside a department, **b**) lowest centre of a commune inside a department, **c**) altitude of territories, **d**) altitude of centres of communes. In the case of a mountain commune like Abriès the mean altitude is unsurprisingly higher than this of its centre. Vertical lines topped with a little squares correspond to edifices inside each commune, indicating each edifice’s altitude. For non-localized edifices both lines and

squares are white, and aligned with the altitude of their commune’s centre).

Some striking observations emerge, that sometimes obviously deny false beliefs. For instance the “classic” vision of Provence’s hilltop villages that dominate the land below appears scarcely in line with reality. In the alpine valleys climate and accessibility constraints, along with the summering activities, lead to impressive densities of edifices in territories deemed as non-wealthy. In Figure 11 the commune of Abriès is underlined - almost all edifices are higher than the commune’s centre, with a rather regular layering that is typical of mountain territories inhabited for a long time span.

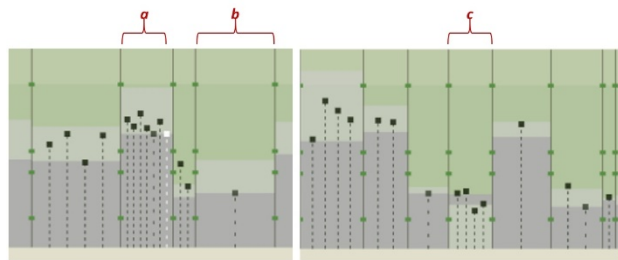


Fig. 12 Left, note the dramatic difference in terms of density between (a) Beauvezer and (b) Bevons – although the latter is lower in altitude, and larger too. Right, (c) Lurs, exemplifies a “hilltop village” pattern, with the centre of the commune higher than its mean altitude.

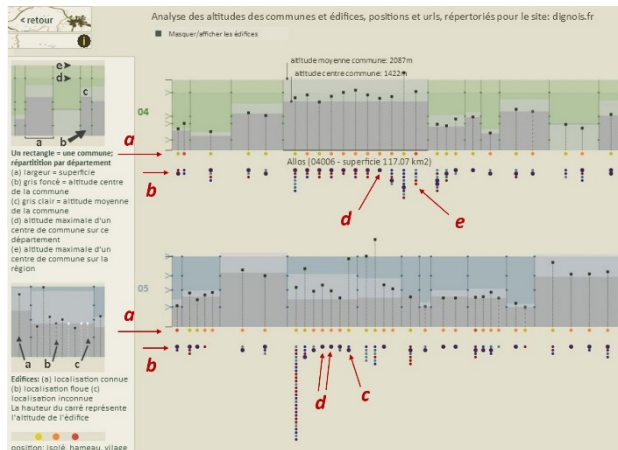


Fig. 13 A version of the densities and altitudes visualisation customized for each information provider (produced on the fly, upon selection of a given information provider, here *dignois.fr*). Only those communes for which the information provider does provide e-content are represented. This version of the visualisation includes indications on the position of the edifice (**a** – isolated vs. hamlet vs. village). It also shows information about e-sources recorded for each edifice (**b** – colours of circles correspond to the type of information providers, see Table 1). The e-sources produced by the information provider are represented by a larger circle (**c**). The visualisation can be used for instance to spot inside a given commune

edifices that the information provider is the only one to document (**d**), or by contrast edifices that the information provider does not document.

4.3 The “orientation cloud”

Christian churches in Europe are, in theory, oriented edifices at least until the Baroque Period (the apse should face the East). Architectural treaties teach us how middle ages builders managed to apply that rule using kind-of basic gnomons. But was that rule really applied to small chapels? Until when? Does the relief or the altitude impact its application? When the rule was not applied, did builders orient edifices erratically? In this visualisation we correlate the orientation with three variables: container (department), date (time anchors), and altitude. Each rectangle corresponds to an edifice, positioned around a “compass”: edifices with an apse facing the north are positioned on the “north” of the compass. The visualisation underlines for instance a tendency of older edifices to be better in line with the rule (Fig. 14). It also shows an inclination of later builders for turning the apse towards the north, and de-correlates the orientation and altitude parameters (in other words, denies the common sense belief that because of stronger relief constraints builders are more keen to step out of the rule).

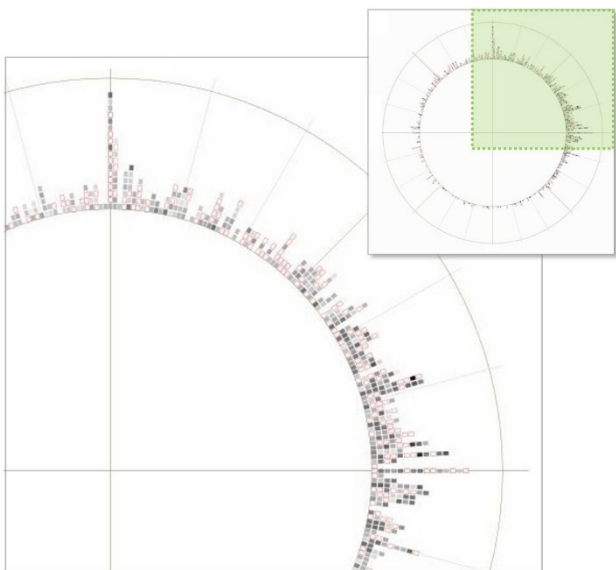


Fig. 14 Orientation cloud (partial view showing the most dense part of the “compass”, the North-East sector) – the darker the rectangle, the older it is. Outlined in red: non dated edifices.

4.4 The “data quality stove”

One of the important observations we wanted to make was to try and spot information patterns in order to gain a better

awareness of the heritage items’ documentation consistency. The following visualisation aims at answering to a simple question: does the quality of the data vary depending on the edifice or the commune considered?

Edifices are grouped by communes, and for each edifice three points aligned vertically convey an indication on (from top to bottom) the localisation, the orientation, and the dating (Fig. 15).

Noticeably what is shown here are not values of these parameters, but a basic yet striking indication as far as our level of analysis is concerned: (a) there is an information available, (b) there are contradictory pieces of information available or (c) no information could be found.

Figure 15 demonstrates one of the uses of the visualisation: getting a global vision with all three parameters present. It is also possible to focus on one and only one parameter and to filter the visualisation’s content accordingly (Fig. 16).

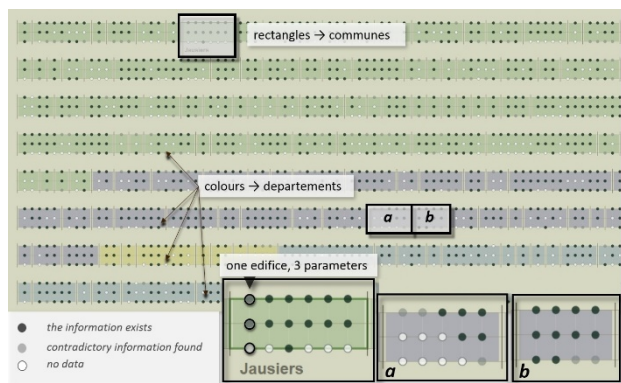


Fig. 15 Components of the quality stove visualisation: three vertical dots correspond to one edifice, edifices are grouped by communes. Communes are associated to a department (background colour). The filling of the dots conveys information on the availability of pieces of information, and on potential contradictions. Note here the significant contrast between communes a and b. Seven out of fifteen parameters recorded as “no data” for commune a, information available for all parameters in commune b, with only in two cases (dating) contradictory information recorded.

Other filtering solutions can be used that give a sense of the challenge ahead if wanting to reach a level of consistency in the documentation of minor heritage assets. In Figure 17 we spot all the communes in which all the edifices are properly documented – their number talks by itself.

The approach is implemented as a web platform combining classic components: an RDBMS (MySQL), a web Server (Apache), a large number of scripts (Perl /PHP) used to produce (on the fly and/or as files depending on

the need) textual content : HTML files, JavaScript variables and arrays, csv “raw data” tables, SVG (visualisations) or geojson data sets.

The interaction between those components is operated through JavaScript components developed on purpose. The cartography bases on the leaflet library [27].

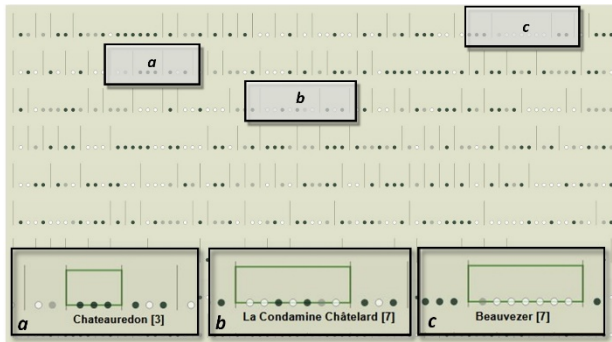


Fig. 16 Focusing on the dating parameter: here three contrasting situations are highlighted, with (a) a commune in which all three edifices in the commune are dated, (c) a commune in which not a single edifice is properly dated, (b) a commune in which all possible cases occur.

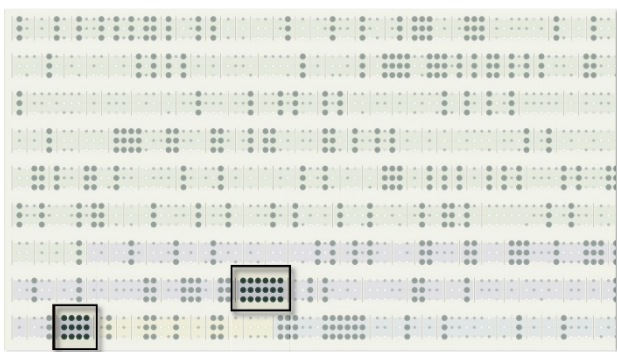


Fig. 17 Highlighting of edifices for which all three parameters are known unambiguously (three large black circles aligned vertically). Not only are these situations a minority, but in only two communes do they appear systematically for all the edifices in the commune. Note that this is a partial view of the visualisation, covering approximately a third of the whole collection. It has to be said, in addition, that the fact all parameters are known does not imply the information recorded is valid and trustworthy: it only says that at least there is an indication available.

5 Evaluation

On the overall thirteen visualisations have been produced up to now, with three of them “customised” for each Information Provider (Fig. 2, 13, 18). Their evaluation has been carried out in a twofold manner: an assessment of their *usability* and *understandability* for non-experts (no

familiarity with InfoVis practices nor with minor heritage), and an assessment of added-value for Information Providers themselves, carried out during ad-hoc workshops. During those workshops, we also collected a more general feedback on the research’s fundamental approach, a feedback we discuss at the end of this section.

5.1 Usability and understandability

The objective of this first round of evaluation was to get a feedback on the graphic choices (encoding, layout, interaction, support for users), but also on the efficiency of the visualisations, *i.e.* on whether or not they do support *information discovery*. To do so six non-experts were asked to fill in one form per visualisation, with each form structured according to the following set of criteria:

- readability assessment (measuring to which extent graphic elements – colours, shapes, *etc.*- are easy to identify, and to differentiate from one another);
- problem accuracy (feedback on parameters that users think should be taken into consideration) [28];
- knowledge communication (Evaluation of the capacities of the visualisation to help users understand pieces of knowledge or to help knowledge holders transfer their knowledge [29]);
- reasoning and hypothesis generation (Evaluation of how the visualisation supports hypothesis generation and interactive cross-examination of data [29]);
- adoption and reuse (feedback on the potential dissemination constraints) [29];
- user guidance (measuring to which extent means to guide the users are relevant and efficient) [30];
- legibility (Verification of the lexical characteristics of the information) [31];
- adaptability (capacity of a system to behave contextually and according to the users’ needs and preferences);
- consistency (the way interface design choices (codes, naming, formats, procedures, *etc.*) are maintained in similar contexts, and are different when applied to different contexts).

We presented to the testers the project as a whole, the data sets and the general relations between pieces of data (*e.g.*, *edifices* located in *communes*, documented by *web sources* that are classified according to *types of info providers*, *etc.*). The evaluation was then performed by each tester on his own, under no supervision: the visualisations themselves were not explained, and testers had to analyse them basing on what they were told of the underlying data,

and on the legends. The actual scenario included on one hand Google forms with series of pre-written questions and on the other hand the online visualisations with as help only their legends.

Answers we collected through that evaluation helped us rethink some basic and early choices such as colour palettes, sizes of graphic elements, distances between elements, level of detail of the legends, or the integration of graduated scales in order to help reading quantities.

We nevertheless acknowledge the fact that such an evaluation, given the small number of testers, and the absence of a trial and error process, should not be considered as more than a starting point. The fact that testers had no background knowledge or involvement in heritage sciences at large leads to an inconclusive evaluation on factors like “problem accuracy”.

5.2 Added-value for Information providers

As researchers, we have made choices in the identification and modelling of the information, and then developed a series of visualisations thanks to which we could perform reasoning tasks and formalize an analytic discourse. But, after all, is all that effort of any help to information providers? Since the information providers we have called in are unfamiliar with information visualisation solutions, will these solutions be of any use for them? And if they are not, are we really into something that has to do with citizen science?

In a paper introducing visual analytics, [32] coined the key expectations that need to be met: understand and analyse our data, understand and analyse our analyses. It was therefore important for us to get direct feedback from information providers on the visualisations. To do so we organized a series of workshops during which we privileged open discussions on each of the visualisations, including the customized ones.

It turned out that they do provide “food for thinking” services on such aspects as density patterns or information quality patterns, and help spot “errors” (typically contradictory data, misinterpretations of temporal hints, *etc.*). But beyond that, and somehow beyond our expectations, the visualisations brought a significant support to workgroup exchanges, for instance on temporal distribution patterns as they emerge from the works of different IPs (Fig. 18). Ultimately nine out of the twelve IPs included in this second round of evaluation expressed the will to get involved in carrying out more such experiments with us, which is some-how serendipitous outcome (as well as a fulfilling one).

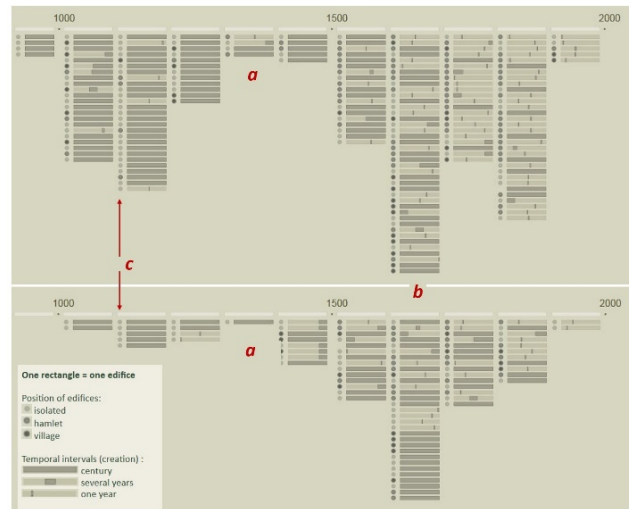


Fig. 18 The customised chronology visualisations for two information providers (top: dignois.fr; bottom: montnice.fr). Each rectangle corresponds to one edifice, the diagram is read from left to right, with each column corresponding to occurrences inside a century (partial views of the visualisations, period prior to 900 AC not shown). Bottom left, legend: graphic encoding of parameters “position of the edifice” and “period of creation”. The data sets correspond to individual actors engaged in the study of two strongly contrasting territories inside the region: *dignois.fr* focuses on one of the least populated French department, and one of the poorest; whereas *montnice.fr* focuses on the hinterland of the Nice - Côte d’Azur territory, a more densely populated area, relatively recently integrated to the French Republic, and not one of the poorest French departments. Note, despite those differences, strong similarities in **a**) - a significant rupture in the creation activity during the 14th century (plague epidemics are a potential, however still unproven, explanatory factor), in **b**) - an intense creation activity during the 17th. In **c**) - an observation that can be made on both data sets: old edifices are in majority isolated edifices – a common sense assertion (no need to erase an edifice with no neighbours...) here backed up by facts. But beyond similarities between the features of the data sets the visualisations underline significant differences: note for instance the proportions of “prior to 1500” edifices in both cases.

The next sub-section sums up remarks we collected during the workshops on more general issues, remarks that go beyond the somehow self-absorbed vision that a short-term research programme encourages. As data analysts and InfoVis solutions designers we naturally tend to consider research products as an end. But this research is not only about showing measurable short term results for our own satisfaction or legitimacy, it is also – and maybe primarily- about circumscribing the motivations, practices, difficulties and challenges ahead for those people who get

engaged in documenting minor heritage items at their own volition and cost. These are the people on which we rely in the long run if we are, as scientists, to foster a better understanding of such heritage assets: noting down and taking into account what they have to say is, from our point of view, definitely not anecdotal.

Furthermore, it is important to state that given the relatively low visibility and public recognition of the importance of minor heritage assets, the amount of actors engaged is limited – we are here very far from the Big Data paradigm. Not only are contributors rather scarce, but there is definitely a clear activity bias in the data we handle, as shown in Fig. 19.

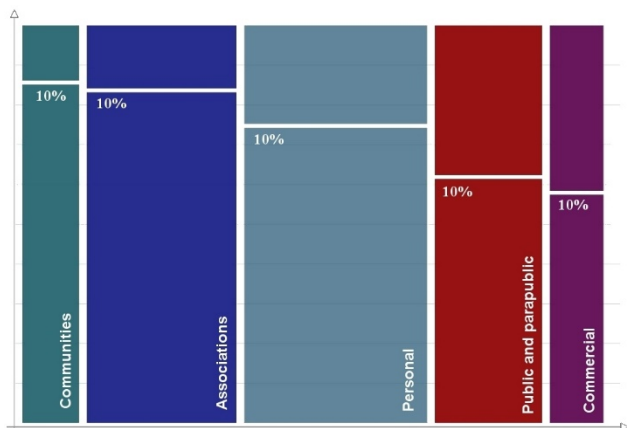


Fig. 19 A basic mosaic display showing the influence of the 10% most active contributors (bottom rectangles) inside each category of IPs. On the x axis, relative importance of each of the five categories of IPs (in number of e-sources quoted). On the y axis, relative importance of the 10% most active contributors inside a category (counted by extracting the root URL of each e-source). Note for instance that one third of the information is available through personal web sites (middle rectangle). In that specific category of IPs, 75% of the information is available thanks to 10% of the contributors (*i.e.* that makes 21 contributors).

Key influential actors here do not number in the thousands, or even in the hundreds, but in the dozens: hence a clear necessity to get a robust feedback, in particular on the sustainability of the information they publish and wish to share.

5.3 Added-value for users

The pieces of information and knowledge that we have harvested are structured according to four main families of descriptors, corresponding to four questions: *where* (geoinformation), *when* (temporal hints), *what* (thematic layers) and *who* (Information providers).

The system provides a number of easy-to-use “classic” query modalities, building on cartographies with different

granularities, chronographies, thumbnails, alphabetically ordered lists, *etc.* Items are associated with external URLs (e-sources) that can be browsed by users from any of the querying modules. In addition, end-user services have been included, with some targeting specifically the information providers (see section 5.2) and some targeting more generally the wide public. The system gives access to raw data such as CSV lists combining various data layers (names, locations, dates, e-sources, *etc.*) or GPX files that can be reused in other contexts. The visualisations that were produced are also available online, but obviously their role is more restricted to analytical tasks, usually carried out by experts rather than by regular users.

As far as acceptability and potential usefulness for such users are concerned, this experience showed the main difficulty we are facing is not the organisation and readability of the information – the interfacing is rather straightforward. The main difficulty is rather having the initiative publicised, and this is an issue that goes beyond this initiative’s field of concern.

More generally evaluating the usefulness of the system’s components (from basic data sheets to visualisations) was something we did want to do, hence the in-depth discussions with IPs as described in section 5.4. We considered that these actors, although they are interested in minor heritage and do invest time on producing information, represent “ordinary people” and could give us a significant feedback. During the workshops we questioned the participants on the usability and usefulness of most of the system’s components using structured forms. We then processed the forms and went back to them in order to discuss potential changes to be made to the system. However we acknowledge that such an effort remains to be done with actors that would not be particularly interested in minor heritage collections. But such an effort would be somehow inconsistent: the actual usefulness of a system, or of pieces of knowledge, can only be judged by those who do need that system, or who do want to acquire new knowledge.

In order to further investigate the issue of helping users what we could plan to do is to try and widen our evaluation effort through focus groups (networking with local actors in the cultural and educational fields) and through web communities (typically Wikidata projects).

5.4 Feedback from Information Providers

The workshops we organized included two different modalities: a rather formal session during which IPs were asked to answer to a predefined series of questions ranging

from their ways of doing and their expectations with regards to data sharing to their evaluation of the platform we have developed, and a series of informal discussions thanks to which we could list a number of basic concerns that they have in common.

Results of formal sessions can be summed up as follows:

- IPs are for most of them engaged in individual initiatives, and decide for themselves on the selection of items they consider for documenting, and on the way they document them (in terms of descriptors, and in terms formatting).
- They are dedicated to have the content they produce made visible for a large community – hence their choice of publishing it on the web.
- They are however sceptical about content published on large community-based platforms such as Wikipedia, and remain committed to content published by an author or an editorial team that takes responsibility for it.
- Yet they are willing to cross-reference their works and eager to enhance their content’s visibility.
- More or less one IP out of two is engaged in publishing content through other means than the web (paper typically), or engaged in local events in and around minor heritage.
- Their feedback on the various contribution modalities proposed is rather ambiguous and hard to build on: no solution emerges as primary.
- They are overwhelmingly in favour of publishing and sharing raw data, yet their understanding of the constraints behind the idea (in particular in terms of standardization and consistency) is rather weak.

Shortly said, what emerges is a pattern combining key priorities: personal commitment, autonomy in all aspects of the publishing process, eagerness to publish and share information on a large scale, with a high visibility, scepticism with regards to community-based un-signed content.

In addition to these results, informal discussions opened a series of unforeseen exchanges and debates. What we view as a potential contribution to a scientific discussion is probably concerns IPs have with regards to traceability, sustainability and open science issues. The following list highlights some of the most prominent concerns we could identify:

- A number of IPs could publish only a part of the material they have gathered over time, and express concerns about their capacity to make use of the rest. They are in need of solutions to share raw data, yet at the

same time are anxious to ensure a clear traceability of the material they could give access to.

- IPs are often given bibliographic or iconographic material in the course of their work, through contacts with other heritage enthusiasts that have not invested on web publishing, or simply by locals who share personal archives with them. Having such material at hand raises for IPs non-trivial questions on what to do with it because of concerns about traceability, licensing and digital rights.
- A major concern in particular for individual IPs is sustainability. Said plainly: “what will become of my work if I stop paying my Internet provider, or when I pass away”. There is no easy answer here, and counting on public or community-based Internet archives is definitely not seen as enough a solution by IPs themselves.

In relation with the previous point, but on a more general level, we witnessed a real concern for the volatility of web-based content. There are lots of reasons why an e-source may become inaccessible: a site may have been closed down for good, a technological evolution from basic HTML to CMS-powered content may have resulted in addressing errors, temporary maintenance works on a provider’s site may result in broken links, *etc.* Figure 20 reuses the grid of e-sources by category of actors presented in Figure 6 but this time with a black square for each e-source unavailable (broken links on Jan 12th 2019). One of the points to be made here is that broken links appear in each and every category of IP. In other words volatility does not only affect individuals who produce content basing on their sole commitment to heritage items, it affects each and every category of IP.

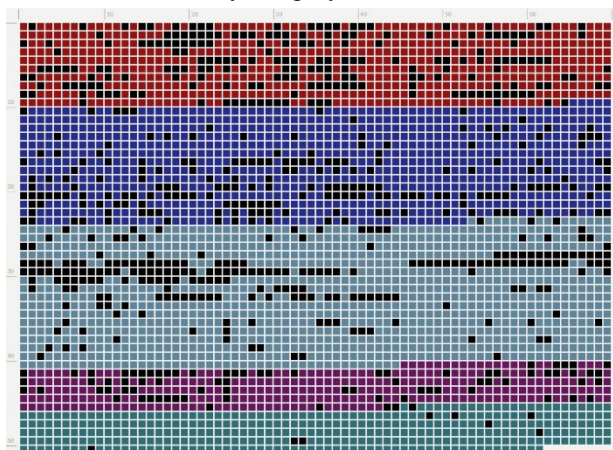


Fig. 20 The distribution of broken links with regards to Information Providers’ profiles. Each square corresponds to an e-source and is coloured according to the type of IP (see Table 1). Black squares correspond to broken links at the time

of query (produced on Jan 12th 2019 – 711 broken links out of 3562).

Briefly said, although IPs do view the web as the place to be in terms of visibility, they express deep concerns on its trustworthiness in terms of “intergenerational” transmission and are left with the paradox of being committed to heritage, but with a sense that what they do may not be transmissible to future generations.

6 Limitations and conclusions

In a recent contribution to JDSA, A. Siebes introduces the idea that we are witnessing a general move from “digitisation” to “datafication”, and pinpoints consequences that data analysts need to anticipate: “*Storing, manipulating and analyzing vast amounts of data of a bewildering variety of types are becoming the core of many new approaches to science, any kind of science*” [33].

Our research can be seen as an exemplification of that move, with data harvesting and analysis tasks strongly impacted by the variety of IPs and of collections. It addresses some pending issues at the intersection of minor heritage preservation, citizen contributions, and spatio-temporal data analysis. The paper does not tell the success story of some new data processing chain, of some new machine learning algorithm that would support for instance cultural tourism actors in their effort to enhance the visibility and comprehensibility of minor heritage assets. It rather tells the intriguing and unfinished story of scarce, unreliable, scattered pieces of data that lose substance and meaning if not regarded as such, and of information providers the motives and practices of whom need to be better understood if we ever aim some day at writing the above mentioned success story.

The paper focuses on three challenging aspects for academics who would want to build on citizen-birthed information sets in order to better document and analyse minor heritage items:

- The heterogeneity of the Information (in terms of scope, of editorial choices, of quality, *etc.*), and behind it the heterogeneity of the Information Providers themselves. We present and discuss the strategy adopted in order to demonstrate potential added-values of such a research on both sides (academics as well as information providers), and in order to pinpoint profiles of information providers.
- The factors of imperfection that are likely to be met when handling such information sets. We first give a global view of the data and information harvested from

citizen-birthed e-sources, and then propose an exemplified list of the key factors of imperfection we came across during the research up to now.

- The design and implementation of visual solutions supporting analytical tasks in the specific context of imperfect information sets. We present and discuss the learnings of some of the solutions we have developed, solutions that are today available online (*territoire.map.cnrs.fr*).

The case study acts as a test bench helping to investigate data harvesting and visualisation challenges. It shows that there is still a significant effort to make in adapting contribution modalities to heterogeneous collections and to the nature of the information we target. It also shows that investing time on the visualisation step, despite information imperfections, is sound: significant patterns emerge. But these patterns are not assertions: *what they renew is our capacity to question and challenge our own level of knowledge and of understanding of those collections*.

Some clear limitations should be quoted, though, at this stage of our research. Making a sound and grounded assessment of the initiative’s added value is arduous since its impact depends on a time taking effort to call contributors in, and to analyse feedbacks. We therefore make no claim that we can present definitive conclusions, but on the overall we believe the experiment shows there is before us a shift in the way academics and collection holders can decode, re-read, augment minor heritage data sets: a shift from “one shot, one collection” protocols to “comparative, cumulative, open science” investigation modalities.

Acknowledgements

This research is funded by the Région Provence-Alpes-Côte d’Azur regional authorities, and conducted in cooperation with the Mucem (Musée des Civilisations de l’Europe et de la Méditerranée) - authors thank E. De Laubrie and Y. Padilla.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest

References

1. J.Y. Blaise, I. Dudek, G. Saygi, Citizen contributions and minor heritage: feedback on modelling and visualising an information mashup. In Proc. DSAA 2018, IEEE 5th International Conference on Data Science and Advanced Analytics, DOI 10.1109/DSAA.2018.00013 (2018)

2. W. Kienreich, "Information and knowledge visualisation: an oblique view". *MiaJournal*, vol.0, No.1 (2006)
3. D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann (Eds), "Mastering the Information Age. Solving Problems with Visual Analytics". Eurographics Association, 2010, <http://diglib.org> (retrieved 21 04 2018) (2018)
4. G. Christoforidis, P. Kefalas, A.N. Papadopoulos, Y. Manlopoulos, Recommendation of Points-of-Interest using Graph Embeddings, In Proc. DSAA 2018, IEEE 5th International Conference on Data Science and Advanced Analytics, DOI 10.1109/DSAA.2018.00013 (2018)
5. C. Quadri, M. Zignani, S. Gaito, G.P. Rossi, On non-routine places in urban human mobility, In Proc. DSAA 2018, IEEE 5th International Conference on Data Science and Advanced Analytics, DOI 10.1109/DSAA.2018.00013 (2018)
6. N. Gershon, "Visualization of an imperfect world", in *IEEE Computer Graphics and Applications*, 18(4), pp. 43–45 (1998)
7. K. Koszewski, "Visualization of Heritage-related Knowledge – Case Study of Graphic Representation of Polish National Inventory of Monuments in Spatial Information Systems", in *Envisioning Architecture: Image, Perception and Communication of Heritage/Kepeczynska-Walczak Anetta* (red.), Lodz University of Technology, pp. 377-387 (2015)
8. D. Myers, A. Dalgity, I. Avramides, "The Arches heritage inventory and management system: a platform for the heritage field" in *Journal of Cultural Heritage Management and Sustainable Development* Vol. 6 No. 2, 213-224 Emerald Group Publishing (2016)
9. P. Le Boeuf, M. Doerr, C.E. Ore, and S. Stead, (Eds), "Definition of the CIDOC Conceptual Reference Model" version 6.2.3 (2018)
10. A. J. Jokar, A. Zipf, P. Mooney, and M. Helbich, *OpenStreetMap in GIScience: Experiences, research, and applications*. Cham: Springer (2015)
11. M.F. Goodchild, "Citizens as sensors: web 2.0 and the volunteering of geographic information", *GeoJournal*, vol. 69, pp.211-221 (2007)
12. L. See, P. Mooney, G.M. Foody, L. Bastin et al., "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information". *ISPRS International Journal of Geo-Information*, 5 (5), 55/1-55/23. ISSN 22209964 (2016)
13. P. Gautreau, M. Noucher, "Sharing Platform in Digital Geographic Information: Everything It Promise?" *Justice Spat./Spat. Justice*, 10 (2016)
14. M. Haklay, "How good is volunteered geographical information? A comparative study of OpenStreetMap & Ordnance Survey datasets". *Environment & Planning B: Planning & Design*, 34, 4, 682703 (2010)
15. S. Spyrtos, M. Lutz and F. Pantisano, "Characteristics of Citizen contributed Geographic Information", Huerta, Schade, Granell (Eds): *Connecting a Digital Europe through Location and Place*. Proc. of the AGILE'2014 International Conference on Geographic Information Science, Castellón (2014)
16. M. Ridge, ed., "Crowdsourcing our cultural heritage", Ashgate Publishing, Ltd. (2014)
17. J. Noordegraaf, A. Bartholomew and A. Eveleigh, "Modelling crowdsourcing for cultural heritage", *Museums and the Web: selected papers from an international conference*, Silver Spring, MD: Museums and the Web LLC, 25-37 (2014)
18. A. Keinanm A. MicroPasts. "An Experiment in Crowdsourcing and Crowdfunding Archaeology". *British Archaeology*, N.139, 50-55 (2014)
19. D. Freeman and J. Freeman, "Use your head: the inside track on the way we think". London: John Murray (2010)
20. A. Wiggins and K. Crowston, "From Conservation to Crowdsourcing: A Typology of Citizen Science", 44th Hawaii International Conference on System Sciences (HICSS), Kauai, HI, pp. 1-10 (2011)
21. W. Aigner, S. Miksch, H. Schumann, C. Tominski, "Visualization of Time-Oriented Data". Springer: Human-Computer Interaction Series (2011)
22. M. Skeels, B. Lee, G. Smith, G. Robertson. "Revealing uncertainty for information visualization", Macmillan Publishers Ltd. 1473-8716 *Information Visualization* Vol. 9, 1, 70–81, [on-line] ;www.palgrave-journals.com/ivs/ (2010)
23. J.Y Blaise and I. Dudek, "Picturing What Others Know: towards a Dashboard for Interdisciplinarity". In Proc. of the 14th I-Know International Conference, pp.15:1–15:8, New York, NY USA: ACM. (2014)
24. Thomson J., Hetzler B., MacEachren A., Gahegan M., and Pavel M., "Typology for Visualizing Uncertainty", In Proc. of the SPIE-VDA 2005: SPIE/IS&T, (Conference on Visualization and Data Analysis), 16-20 January 2005, San Jose, CA USA (2005)
25. Zuk T. and Carpendale S., "Visualization of Uncertainty and Reasoning", [in] A. Butz et al. (Eds.): *SG 2007, LNCS 4569*, Springer-Verlag Berlin Heidelberg, pp. 164–177 (2007)
26. Friendly, M., "Visions and Re-visions of Charles Joseph Minard", *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 1, pp. 31-51. (2002)
27. Leaflet open-source JavaScript library for mobile-friendly interactive maps. <http://leafletjs.com/> (2018)
28. T. Munzner., "Visualization Analysis and Design". AK Peters Visualization Series, CRC Press (2014)
29. H. Lam, E. Bertini, P. Isenberg, C. Plaisant, S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios". *IEEE Transactions on Visualization and Computer Graphics*, Institute of Electrical and Electronics Engineers, 18 (9), pp.1520– 1536, (2012)
30. S. Beier, "Reading Letters: designing for legibility", BIS Publishers, pp.1-190 (2012)
31. J. M. Christian Bastien and Dominique L. Scapin, "Ergonomic Criteria for the Evaluation of Human-Computer Interfaces Technical report", N. 156 INRIA (1993)
32. D. Keim, G. Andrienko, J.D. Fekete, C. Görg, J. Kohlhammer, "Visual Analytics: Definition, Process and Challenges".

Information Visualization - Human-Centered Issues and Perspectives, Springer, pp. 154-175, LNCS (2008)

33. A. Siebes, "Data science as a language: challenges for computer science—a position paper". *International Journal of Data Science and Analytics*, Vol 6, pp.177–187 <https://doi.org/10.1007/s41060-018-0103-4> (2018)