inDICEs

Measuring the Impact of Digital Culture

Deliverable 1.3

Report on data gathering v. 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870792.

The sole responsibility for the content of this publication lies with the authors. It does not necessarily represent the opinion of the European Union. Neither the EASME nor the European Commission is responsible for any use that may be made of the information contained therein.

The research leading to these results has received funding from the European Union's Horizon 2020 Programme (H2020-DT-GOVERNANCE-13-2019) under grant agreement n° 870792.

inDICEs

Measuring the Impact of Digital Culture

D1.3 Report on data gathering v1

Final version

Grant Agreement number:	**870792**
Project acronym:	**inDICEs**
Project title:	**Measuring the impact of Digital CulturE**
Funding Scheme:	**H2020-DT-GOVERNANCE-13-2019**
Project co-ordinator name, Title and Organisation:	**Simonetta Buttò, Director of the Central Institute for the Union Catalogue of the Italian Libraries (ICCU)**
Tel:	**+39 06 49210425**
E-mail:	**simonetta.butto@beniculturali.it**
Project website address:	**http://indices-culture.eu/**

Author:

Pier Luigi Sacco, FBK

Manlio De Domenico, FBK

Oriol Artime, FBK

Federico Pilati, IULM University

Maria Tartari, IULM University

Partners:

Katinka Böhm, WLT

Alba Irollo, Europeana Foundation

Franco Niccolucci, PIN

Sonsoles Parajes, KU Leuven

Rasa Bocyte, NISV

Sara Uboldi, University of Modena and Reggio Emilia

Sabrina Pedrini, Università Alma Mater Studiorum Bologna

Document History

- 15.09.2020 – Draft Version
- 16.11.2020 – Draft Version
- 01.12.2020 – Draft Version
- 11.12.2020 – Draft Version
- 11.12.2020 - FIRST SUBMISSION TO PARTNERS
- 04.01.2021 - internal review
- 09.01.2021 - internal review
- 12.01.2021 - internal review
- 09.02.2021 - internal review
- 11. 02. 2021 - internal review
- 24. 02. 2021 - SECOND SUBMISSION TO PARTNERS
- 05. 03. 2021 - second internal review
- 18. 03. 2021 - second internal review
- 29. 03. 2021 - FINAL VERSION

Table of Contents

1. Executive Summary

This Deliverable aims to briefly describe the data collection processes, the datasets gathered and the preliminary data analysis on users' behavioural changes that was carried out by the WP1 working group.

The inDICEs data collection processed and/or stored within the first 12 months of the project consists of

a) data analyzed as part of the inDICEs participatory platform, where results are made available through the Open Observatory

b) data of relevance provided by third-parties such as:

- Enumerate
- Nemo
- Eurostat
- State of the commons
- United Nations Conference on Trade and Development
- Digital Economy and Society Index
- EU open data portal

c) online content gathered continuously, made accessible by means of the Visual Analytics Dashboard that covers:

- Online news and web sources
- Twitter posts
- Youtube videos
- Facebook pages

d) FBK collected on-line datasets on cultural production, from the following sources:

- Wikipedia
- Tiktok
- Deviantart
- AllTheater
- IMDB

and was gathered with the purpose to:

- monitor and analyze the state of cultural digitization via WLT analytical tools and through the Visual Analytical Dashboard, configured for culture-based web sources (news, websites, social networks, blogs, forums) and with domain-relevant keywords according to a series of pre-sets and new indicators [as described in Deliverable D1.1].

- stimulate behavioral changes in the users of participatory platforms in order to favor production and access. To understand how this process of collective cultural production works, inDICEs chose Wikipedia as its first case study, in order to extract new useful indicators to fill the open Repository available for single researchers and institutes.

2. Introduction and Objectives

The objective of this deliverable is to describe the work processes related to the strategies of data gathering and analysis employed during this first year of activity [M1-12] by the WP1 working group and by all consortium partners; to describe the results achieved by this work package, as well as to briefly present the datasets and give an overview of the preliminary data analysis conducted on the behavioural patterns of users.

This is the first of four data gathering periodic reports that outlines the status of the first phase of the inDICEs project, namely the data gathering activity, with specific information on the quality, reliability and accessibility of the gathered information. It also describes the data collection process of the different datasets that were gathered, processed and/or stored within the first 12 months of the project, the limitations that emerged, and the plans for the next 6 months [M15-21] that will be aimed at filling the most relevant gaps and strengthening the sustainability of the project.

The data gathering strategy has been devised, for one thing, to collect brand new data on cultural digital platforms' users behavior, and on the other, to conduct complementary analysis to the ones made possible by already available on-line sources and relevant reports on CHI digitization status and socio-economic impact. In particular, strategies of acquisition of relevant data through social media and from digital platforms of interest from the European culturescape have been defined and implemented.

Also, these resources are likely to be useful for a wide range of researchers and practitioners in cultural and creative sectors, not just for users interested in the digital humanities sector, in line with the targeted profiles identified by WP4.

To provide a useful and flexible report for inDICEs users, we have developed at first a literature meta analysis on the already existing wealth of data gathered from statistical institutes and other data gathering institutions, which will be available in the inDICEs Repository (see chapter 7.1). inDICEs collected, aligned and integrated the already existing wealth of data gathered from such institutes and institutions, as well as from the scientific literature on the topic, in addition to the data contained in available archival sources relative to the effects of the digital revolution on the cultural and creative ecosystem.

The data collected will be fed into the Open Observatory infrastructure in the InDICEs portal as detailed in Deliverable D4.1, and into an additional component that will be developed within the next few months, as detailed in the "Plan for the next period section", namely the Repository.

3. Literature meta-analysis on already existing wealth of data gathered from statistical institutes and other data gathering institutions

The aim of this meta-analysis is to provide a useful and flexible document which summarizes the already existing wealth of data on CHI digitization gathered from the most important statistical (or other) institutions.

3.1 ENUMERATE 4.4: state of the art on data gathering on CHI digitalisation

Enumerate is the project by which inDICEs was inspired. Therefore, it constituted a starting point for collecting statistical data about digitization, digital preservation and online access to cultural heritage in Europe.

The project was funded under the EC's ICT Policy Support Programme, and investigates the state of digitisation in cultural heritage institutions in Europe, particularly museums, libraries, archives. It aims to provide a baseline of data that can inform decisions at the national and EU policy level, and is based on gathering statistical information through a network of national coordinators. Since 2011, it has run four surveys.

The last survey was conducted in 2017 and structured around six topics: 1) digital collections; 2) digitisation activity; 3) digital access; 4) participation; 5) digital preservation; 6) digital expenditures. Nearly 1,000 institutions took part in it and 82% of them claimed to have a digital collection or to be in the process of launching a digitisation project.

Europeana has contributed to Enumerate and currently plays a major role in ensuring its legacy. The Europeana PRO website currently hosts the Enumerate Observatory, which provides a reliable baseline of statistical data about digitization, digital preservation and online access to cultural heritage in Europe. Enumerate collects statistics through surveys, reuses data from existing research, analyses and publishes the results, develops indicators and explores CHIs' needs. These resources make the documentation related to the project available, and the datasets of the surveys are anonymised and provided as raw data to users.

Some main findings of the report:

- 82% of the respondents have a digital collection. Most institutes have a rich mix of different cultural heritage materials.
- 42% of the institutions have a written digitisation strategy (was 41% in 2015).
- More than half of the institutions (59%) collect native digital items.

- Overall, institutions report that they have 51% of their descriptive metadata online for general use. Libraries are at the high end for this indicator (76%), whereas museums have the lowest score (33%)
- Academic research is viewed as the most important reason to offer digital access to a collection (8.8 on a 10-point scale), followed by educational use (8.5). Sales and commercial licenses are deemed least important.
- 42% of the digital objects managed by the participating institutes are not available online.
- A notable outcome is that respondents foresee a decline (-4%) in the number of objects digitally available through their own website in the next two years subsequent to the 2017 survey. Respondents do expect an increase via external channels, like Social Media (+25%), Wikipedia (+14%), Europeana (+5%) and other aggregators (+11%).
- 45% of the institutions do not have a solution yet for long term preservation based on international standards for digital preservation (was 47% in 2015).

3.2 Meta analysis and comment on existing wealth of data gathered from statistical institutes and other data gathering institutions

European statistics about cultural heritage (CH) institutions lack a global perspective, and as a result the data is fragmentary.

A distinction must be made between the so-called GLAMs (Galleries, Libraries, Archives and Museums) and Monuments and Sites, indicated henceforth with the acronym M&S. Nothing coherent exists, to the best of our knowledge, for intangible heritage, besides the information provided in the UNESCO world heritage list, which however is not collected in a uniform way. For instance, it is difficult (if not impossible) to evaluate the number of people practicing *Falconry* in the 8 EU countries enlisted for this activity (Austria, Belgium, Czech Republic, France, Hungary, Italy, and Spain); how many herdsmen are involved in *Transhumance* in Austria, Greece and Italy; or how many *Sicilian Puppets Theatres* are still performing in the streets today. Thus, we will not consider this important branch of CH.

As regards M&S, even compiling a list at the European level is made difficult by the different national regulations. For example, in Italy the approach is a binary yes/no: an edifice or a site belongs officially to M&S if it is “notified”, i.e. there is an administrative act by the competent Ministry stating that this item has historical or artistic importance and therefore must be preserved. The declaration may concern the whole building or only a part of it. No public list is available also because of privacy considerations, but some statistics are provided by MIBACT[1]. It is possible to generate reports online; the results are available by

[1] http://vincoliinrete.beniculturali.it/VincoliInRete/vir/statistics/redirectReport3

region. From these, one can find for example that in Tuscany there are 13174 notified buildings, distinguished in finely detailed subcategories, for example “church”, “oratory”, “pieve”, “chapel”, “basilica”, “baptistry”, “collegiate” and so on, with several other categories grouping monastic buildings. No totals are available, and therefore figures must be added up at user level. No data are provided in this summary report about period, style or any other architectural information. Collecting such detailed data would need permission by the Ministry because, as already mentioned, access to details is reserved, and only quantity is publicly available. Considering that there are 20 regions, a rough estimate gives 150,000 - 200,000 notified buildings. In France the system envisages two levels of classification for historic buildings, according to their national or regional importance. About 45,000 monuments are listed. These are just two examples, showing there is a huge variety of classifications systems and data about M&S across Europe. In conclusion, building suitable statistics out of this wealth of data is a research project on its own.

For archaeological heritage, EAC (Europae Archaeologiae Consilium) – the network of heads of national services responsible by law for the management of the archaeological heritage in the Council of Europe member states – is concerned with policies and best practices but does not collect statistical information.

As concerns museums, statistics are available in all EU countries but often organized in various ways. Since they are usually produced by public organizations, they generally focus on state-owned museums which are only a part of the whole picture. They may neither include privately owned ones, among others religious museums, nor civic and regional ones. Most small museums are thus left out, although they in no way are minor ones. Museum associations as ICOM aggregate only a part of the community.

Museum statistics are collected by EGMUS[2]. The collection is based on a questionnaire provided by participating countries, which include all EU countries except for Cyprus and Malta. Unfortunately, questionnaires are updated at different times, so the pieces of information may be outdated and are usually not aligned for all countries. Nevertheless, EGMUS is the best source for museum statistics in Europe. The questionnaires are filled by high-level officers from the contributing countries and sources are referenced. The data they publish is highly reliable. It must be noted that EGMUS collects data also from non-EU countries [for detailed tables obtained processing EGMUS statistics, see Annex 1].

An estimate of the total number of EU museums can be obtained from table 1 (see Annex 1). Considering the lack of answers about non-state-owned museums in Italy and Greece, the total may be estimated to be about 18 000 museums. It must be also noted that the definition

[2] https://www.egmus.eu/nc/en/statistics/complete_data/

of museum is slightly different from country to country. The balance between state-owned or managed museums and the others depends on national regulations – for example, in Germany it depends on the federal state structure that assigns the responsibility of cultural heritage to each Länder. In most countries, the ratio of art & archaeology museums to science & technology is 2:1 or more – note that data from Germany are unavailable because of a different classification system, but in any case a lower ratio should be expected.

Table 2 (see Annex 1) shows a strong penetration of computers in museum environments, at least in the countries that provide the information.
Considering only the countries that answer to this question, this survey gives only partially significant results, because partial information excludes, among others, large countries such as France, Italy and Germany, and includes only 25% of the total number of museums. However, as for the responding countries, 77% of the museums have at least one computer, with several countries reaching 99%. In conclusion, EGMUS provides reliable data, but the outcomes do not cover all aspects. In particular, no information is collected about the presence of more up-to-date IT services such as social networks, or about the way in which on-site IT-based presentations are implemented. It is probable that, even if asked, the contributors to EGMUS have no information to provide in this regard; we suspect that almost nowhere such information is collected at national level.

Eurostat – the statistical office of the European Union, responsible for publishing high-quality Europe-wide statistics and indicators that enable comparisons between countries and regions – collects and publishes data on culture. According to their site[3], such statistics include, with various degrees of systematicity and completeness:

- Cultural employment;
- Characteristics and performance of enterprises engaged in cultural economic activities & sold production of cultural goods;
- International trade in cultural goods;
- International trade in cultural services;
- Cultural participation (practice and attendance) and culture in cities (such as satisfaction with cultural facilities of cities' residents and 'cultural infrastructure');
- Private (household) expenditure on cultural goods and services;
- Price index of cultural goods and services;
- Public (government) expenditure on culture.

Therefore, such statistics do not include any information at all about cultural heritage.

[3] https://ec.europa.eu/eurostat/web/culture

Some figures are occasionally included in the Commission's report[4] "European Commission report on Cultural Heritage: Digitisation, Online Accessibility and Digital Preservation" (latest update 2019), but this report addresses policies rather than data. This activity also produces national reports which occasionally include numbers for specific cases. No global data collection is considered.

Finally, there have been two initiatives collecting data about cultural heritage-related digital activities.

Already mentioned earlier, the first one was carried out in 2017 by the project ENUMERATE and summarized in a report[5]: the investigation was carried out through questionnaires. The survey authors admit that for some countries the response was "disappointing". These countries include France (0 answers), Germany (29 answers), and Spain (26 answers). Also Italy did not perform well (39 answers). Only about 1/3 of the total respondents are from museums, which further reduces the value of the statistics for our case – if such category percentage is projected on responses, for the above-mentioned four countries less than 1% of museums replied to the survey. Moreover, other data indicates that survey participants come from medium to large museums, as the average number of paid staff is about 49 FTE.

It is reported that 77% of the museums have digital collections or are involved in digitization, while 45% have a written digital strategy. Such strategies concern long term preservation (51%), publishing digital collections (77%), acquisition of digital collections (31%) and digitization of analog (print, manuscripts, physical items, etc) collections (91%). Native digital collections are more present in ethnographic museums (62%) and in art museums (58%) than in archaeological museums (46%) and science, technology (37%) or natural science ones (31%). This is not unexpected, as ethnographic museums have more sound recordings than the others, and art museums usually document art works with images.

The following table reports the nature of digital resources for museums. The percentage is the number of institutions having such resources over the total number of museums.

Nature of digital resource	Text based	Visual 2D	Visual 3D	Interactive resources

[4] https://ec.europa.eu/digital-single-market/en/news/european-commission-report-cultural-heritage-digitisation-online-accessibility-and-digital

[5] https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Deliverables/d4.4-report-on-enumerate-core-survey-4.pdf

Percentage of museums with such resource	42%	64%	45%	46%

From this table, it appears that about half of the museums have digital resources. Considering that since the surveyed institutions are mainly large or medium-sized ones and in smaller ones the overall situation might be worse, this suggests that digitization has not yet reached a critical mass.

Another statistic concerns the progress of digitisation. For museums, 31% of the collections have been already digitally reproduced, while 57% still need to undergo digitisation (the remainder is not suitable for digitization). In other words, only one third of the collections has been digitized while two thirds have not. As regards online access to such collections, according to ENUMERATE “almost all institutions provide online access to both the metadata and the collection, but there are still parts of the digital collection that can only be accessed as metadata, or not accessed at all online”. About half of the digital collections are accessible online, and an additional 17% is available on site; the remainder is not accessible by the public. Access to digital collections is mainly via the institutional website or Europeana (about a half), while social platforms account only for 8%. Unfortunately, these figures are not disaggregated for museums. Finally, ENUMERATE provides some information about the number of accesses to digital collections. It is interesting to note that large and very large institutions (with budgets greater than 0.5 million Euro) have 95% of the visits.

In conclusion, ENUMERATE produces an insight into the progress of digitization in museums but the results of the survey are biased by the number and distribution of respondents. Of all respondents for all categories, 16% were from the Netherlands; 15% from Sweden; 12% from Poland; about 6% each from Czechia and Greece; 4% each from Hungary. Latvia and Slovenia. Altogether, these countries account for two thirds of the respondents. On the other hand, the same countries add up to about 10% of European museums. Overall, the responses concerning museums (37% of the total, i.e. 364) amount to 2% of European museums. Moreover, the sample selection was not made according to statistical sampling rules, but depended on the availability, interest and willingness of respondents to collaborate. The fact that most responses came from large institutions is not casual: probably in these institutions there is more available/interested/competent staff than in smaller ones.

A different kind of survey has been recently carried out by the NEMO museums association. NEMO[6] is an organization including national museum bodies, museum networks and associations and individual institutions. Its activities include networking museums, developing and assisting them in developing special projects, and training museum staff.

With the explosion of the COVID-19 pandemic, NEMO at short notice organized a survey about how museums were reacting to the regional emergency measures put in place that caused the closure of most of them to the public. The survey was carried out between 24 March and 30 April 2020.

Table 3 (see Annex 1) lists the number of respondents by country.

The percentage columns contain, respectively, the percentage of the number of answers to the NEMO survey and the percentage of the number of museums. The last column contains the ratio of the two percentages – a number greater than 1 shows that the country influences the result more than its relative weight as to the number of museums, and indicates the imbalance between the country weight in the NEMO survey compared to its weight according to the number of museums.

Table 3 shows that Austria, Belgium, Greece, Latvia, Lithuania, Luxembourg, Slovenia, Sweden and to a lesser extent Denmark and Finland influence the results more than due (Greece because the number of museums provided by EGMUS is largely underestimated); Bulgaria, Belgium, France, Germany, Ireland, Poland and Slovakia less or much less than due according to their number of museums. Among the countries with about or more than 1000 museums, France, Germany and Poland are underrepresented, Italy (considering circa 900 museums, see note above) and Spain have a fair representation. This consideration undermines the statistical value of the survey, which however remains an important source of information about the reaction of museums to the COVID-19 emergency.

Moreover, it should be noted that the survey involved only less than 5% of the total number of museums. This was not a representative sample: it was the result of NEMO's capacity to spread the questionnaire and of the availability/willingness of museum staff to reply, which may depend on several factors, both objective and subjective. A large majority of respondents came from medium-sized museums (75% from museums with 20,000 to 100,000 visitors per annum).

Among the survey results, one may quote the following:

- The presence online was greatly increased during the peak of the pandemic, reportedly in 80% of the institutions, with a preference for social media, possibly

[6] https://www.ne-mo.org/

because they do not require extra investments, time, costs and skills. Indeed, the staff may have greater familiarity with them for personal reasons.

- The response of the public was very positive, with an average increase of visits by 25%, again with a preference for social media, especially for Facebook and Instagram. They are followed in popularity by educational material (possibly for the extended closure of schools), videos and films – easier to consume than virtual visits.

NEMO carried out a second survey between 30 October and 29 November 2020. This focused on the impact of COVID-19 on museums, especially as regards their closure and its permanent effects. Also safety measures have been considered. This survey confirms the indications of the former. It also provides support for policy recommendations, which are the focus of a third report that mainly addresses policy makers and funders.

All in all, the NEMO surveys are useful results of an admirable effort by a small organization that was capable of rapidly analysing the state of affairs in a dramatic situation involving the whole of society, and calling attention on the value of culture even when health risks and economic disruption are rightfully the main concern.

The NEMO reports also provide a concise but effective perspective on the trends regarding the digital future of museums which was previously unavailable, and that goes beyond the emergency situation that prompted the investigation.

3.3 Covid-19 impacts on CCS: best practices in EU

During the last year, the COVID-19 pandemic has reshaped the cultural ecosystem, and there has been an increasing attention among InDICEs consortium partners in considering this unexpected development as a major issue of interest in the future unfolding of the project. The cultural sector, even if its economic sustainability has been seriously disrupted, has rapidly responded through innovative strategies of content production and diffusion, especially in terms of accelerated digitalisation, strategic development of sustainable businesses, and finance models for the cultural and creative (CC) sector. Digital infrastructure is on the verge of a structural revolution, triggered mainly from the bottom-up, in order to improve participation, access, and overcome digital inequalities (e.g. OECD data shows that "regional differences in household access to broadband are significant, with variations between capital regions and other regions reaching over 30 percentage points in some countries").

An increasing access to digital library resources, virtual visits to museums and visual arts exhibitions, as well as the increase of online concerts and theatre, dance and opera performances, reveal the urgent need for culture as a key sensemaking dimension of everyday life. This crisis sparked new awareness of the potential of cultural e-participation on cultural digital platforms. According to the European Commission's JRC report "European Cultural and Creative Cities in COVID-19 times" (2020), *many local governments and cultural institutions have promptly reacted to CCS Covid-19 crisis with grants, financial and non-financial measures for digital re-organization and innovation to support citizen participation to the local cultural scene, mostly about the extremely rich creation of new digital channels, campaigns and portals to offer people opportunities to enjoy culture safely at home*. According to the "OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis", *the 2020's massive digitalisation coupled with emerging technologies, such as virtual and augmented realities or new technologies that allow social aggregation and exchange in virtual communities, is creating new forms of cultural production, dissemination, new business models with market potential, and can impact strongly on cultural ways of unfolding e-participation.*

OECD policy recommendations for cultural and creative sectors in light of COVID-19, in the mid-term, regard investments in digital infrastructure aimed at amplifying advances in cultural and creative sectors. The main benefit can be seen in the long term as education can profit from advances in cultural and creative sectors, particularly in the use of new digital tools that build on gaming technologies and new forms of cultural content.

According to the *Agenda21Culture* report "Culture, Cities and the COVID-19 Pandemic", *the crisis ha produced a multiplication of cultural initiatives, sharing of digital resources, collections, videos, photos, etc…, promoted by public institutions, networks, but also by citizens.*

Here are some best practices experimented to encourage convergence across cultural work, cultural participation and digital potentials by re-thinking cultural offer in EU cities (as reported by: UCLG, 2020; KEA, 2020; OECD, 2020; Montalto et al, 2020):

- Leeds' theatre company https://www.slunglow.org/ appeared regularly online;
- in Turin, the Goddess Factory created a new format for the online streaming of performances[7];
- Barcelona cultural portal has been put at the service of public and private initiatives such as museum, shows, talks, and concerts[8];
- Valencia has created on-line channels and programs to promote exhibitions, events and the sharing of digital books from the municipal cultural institution Biblioteca Valenciana[9];

[7] https://www.livedelivery.it/

[8] https://www.barcelona.cat/barcelonacultura/es

[9] https://www.hel.fi/uutiset/en/kaupunginkanslia/helsinki-offers-express-funding-to-culture-sports-youth-work-school-lunches-to-continue

- the City of Helsinki Network provided a wide variety of e-books, e-magazines, language courses and e-music services in different languages through its shared e-library service[10];
- Warsaw is promoting cultural events produced by the municipal institution that have moved on-line[11];
- Berlin created an on-line platform to connect cultural institutions and the public with discussions, performances, opera concerts and vernissages on-line, supporting artists through online donations[12];
- Rome launched the initiative #laculturaincasa, which includes access and digital resources of libraries, virtual tours of museums, the contest #FinestresuRoma, as well as live online music, theatre and opera performances;
- The city of Bilbao created the initiative "Me quedo en casa" with a special agenda of cultural activities online[13];
- The city of Malmö has launched an online aggregator of performances and exhibitions – theatre, opera, concerts, exhibitions, conversations with authors, tips on art activities for children, among others[14];
- In Paris, 14 city museums and cultural entities are giving free online access to more than 320.000 items, virtual visits of recent exhibitions or special contents for audiences;
- In Dublin, the Culture Company launched activities online, including dance, singing, painting and poetry-writing to virtual online classes, and planned to support cultural connections. Culture Clubs and talks programmes moved online, and Our City Our Books started collecting recommended reads and sharing book suggestions by people who live in Dublin[15];
- Trois Rivieres launched the movement #enmodevirtuel to bring together all the offers under a single address, in order to ensure the continuity of its actions in the cultural milieu;
- Terrassa also reinforced the connection with citizens by involving them in actions in different cultural fields (visual arts, literature, cultural recommendations). The Library Network, the Museum and the Archive of the city organised special activities and

[10] https://www.hel.fi/uutiset/en/kaupunginkanslia/helsinki-offers-express-funding-to-culture-sports-youth-work-school-lunches-to-continue

[11] https://www.um.warszawa.pl/aktualnosci/zosta-w-domu-z-kultur

[12] https://www.berlinalive.de/

[13] https://www.youtube.com/channel/UCMTqxrE7Y3oaMTYCM9iDxRQ

[14] http://www.agenda21culture.net/culture-malmo-covid-19

[15] https://www.dublincitycouncilculturecompany.ie/

different cultural sectors helped to create an online cultural agenda. In general, a greater awareness has raised the importance of online communication and tools;

- Some cities – such as Bologna, Lisbon and Lyon – are participating in the ROCK EU programme, which is based on the promotion of cultural heritage and historic city centres, and have reflected on democratic access and communication as well[16];
- Eindhoven is working to enable online participation of cultural stakeholders in all kinds of projects;
- Vilnius is fostering a new, more constructive work culture based on public e-services, public hearings on architecture and urban development, and educational and training programmes;
- Athens is using the large amount of cultural content which was previously absent from the digital sphere and that has been suddenly made accessible as an opportunity to connect with local communities.

[16] https://rockproject.eu/

4. Data on Economic impact: an analysis of existing statistical data sources

As part of the Methodological Toolbox, a state-of-the-art review of existing statistical data sources is provided to all partners in order to give a smart instrument for being aligned on the current state of CHI digitalization, thanks to the integration and incorporation of data and information currently fragmented in various reports on CHI digitization and socio-economic impact published by organisations or projects: the above-mentioned ENUMERATE, NEMO, EGMUS, and CDSI, EUROSTAT, DESI, UNCTAD, EU Open Data Portal, SotCommons.
All the data here reported will be available in the Repository.

4.1 The Digital Economy and Society Index (DESI) 2020

The Digital Economy and Society[17] Index is a European Commission's yearly index that aims at measuring the digital competitiveness of the EU Member States and its evolution. It is based on several indicators that are grouped in the following areas:

- *Connectivity:* the analysis mainly measures the broadband market developments in the EU.
- *Human capital and digital skills:* digital skills are gaining an increased importance in the EU as not only the digital infrastructure is needed, but also the ability to be digitally proficient for the society to take advantage of all the potential of the digital environment. Therefore, the analysis looks at the barriers preventing EU households from getting internet access and at the digital skills of the EU society (from software skills to more advanced skills that allow working in the digital environment, such as ICT specialists).
- *Use of internet services by citizens:* this thematic area looks at the engagement of citizens that have internet connection and digital skills to engage in online activities. Such activities include the use of online activities, the use of communication activities and the use of online transactions such as shopping or online banking.
- *Integration of digital technology by businesses*: this area examines digitization, namely the integration of new technologies, by businesses and e-commerce.

[17] See: https://ec.europa.eu/digital-single-market/en/digital-economy-and-society-index-desi#:~:text=The%20Digital%20Economy%20and%20Society%20Index%20(DESI)%20is%20a%20composite,Member%20States%20in%20digital%20competitiveness;

- *Digital public services:* the analysis mainly measures the uses of digital technologies in governmental organizations, including the demand and supply of digital public services and open data.
- *Research and development ICT:* different indicators in relation to the ICT sector and the R&D performance are measured in this area, e.g. the added value of the ICT sector or the employment and productivity of the sector. The Commission also publishes a list of ICT projects funded under the H2020 framework programme.

In addition, the international DESI extends the analysis to other 18 non-EU-countries, and the Women in Digital Scoreboard provides an analysis of the female inclusion in digital entrepreneurship, careers and jobs.

The European Commission publishes the data set yearly, including the indicators used and a report for each area of assessment. Despite the fact that digitization of cultural heritage and digital content provided by cultural heritage institutions is not measured in any of the assessed areas, there are certain interesting conclusions that can be highlighted as they may have an impact on the access to digital content online.

First of all, in the area of digital skills, it is interesting to mention that in 2020 there is an improvement in the number of people that are achieving at least basic digital skills (internet skills) accounting to 58% of the population. Yet, there is still a large amount of the EU population lacking such skills. Furthermore, there is a clear shortage of ICT specialists in most of the EU countries. The EU Code Week is an event that brings together volunteers, teachers and digital ambassadors, and is supported by the European commission as part of its Digital Single Market Strategy and the Digital Education Action Plan.

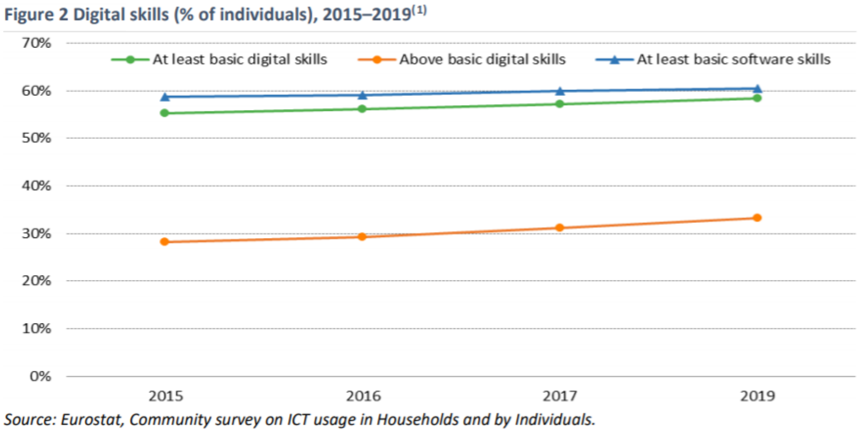


Figure 2 Digital skills (% of individuals), 2015–2019[(1)]

Source: Eurostat, Community survey on ICT usage in Households and by Individuals.

Figure 1 DESI area report: Human Capital/Digital skills 2020[18]

[18] See DESI Report: human capital and digital skills, pg.5

Secondly, in the area of the uses of internet services by citizens, it must be mentioned that one of the analysed areas is the engagement in certain online activities. Within this area, certain key indicators that are interesting for the inDICEs project (although not directly linked to cultural heritage institutions) have been measured, such as the access to music, videos and games, access to videos on demand and access to news online.

The report highlights that using the internet for listening to music, playing games and watching videos are the most frequently chosen activities on the internet (81% of individuals who used internet in the last 3 months) followed by reading the news (72%), shopping (71%) and banking online (66%)[19]. Other indicators measured were the use of video calls (which increased considerably) and the use of social networks.

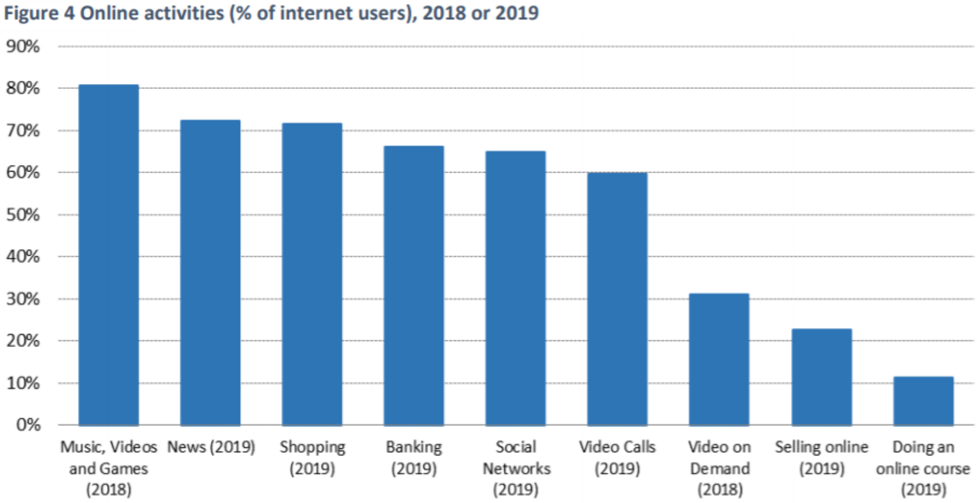


Figure 4 Online activities (% of internet users), 2018 or 2019

Source: Eurostat, Community survey on ICT usage in Households and by Individuals.

Figure 2 DESI Report: uses of internet and online activities[20]

In the area of *e-commerce,* the analysis shows an upward trend in engaging in shopping online in the EU countries, but what is interesting to highlight is that books, magazines and newspapers are one of the most popular categories of online purchases (33%). Films and music were mostly purchased by young people 16-24 years old (34%)[21].

The analysis of the area of digital technologies by enterprises clearly shows that businesses are more and more digitized in the EU. Yet, the adoption of digital technologies is much more common in large enterprises than in SMEs. Few SMEs sell online (18%) and only 8% sell across borders online[22].

[19] See DESI Report: Use of Internet and Online Activities, pg.5

[20] See DESI Report: Use of Internet and Online Activities, pg. 6

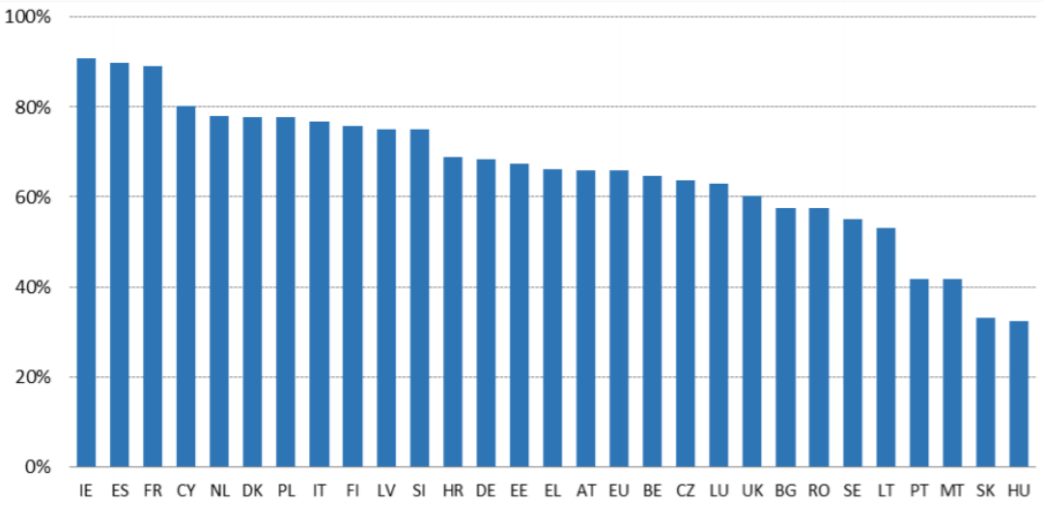
[21] See DESI Report: Uses of internet and online activities, pg. 7

[22] See DESI Report: Integration of Digital Technology, pg.7

In the area of digital public services, the most interesting part for the inDICEs project is the analysis of the use of open data by the EU governmental institutions. In this regard, this indicator measures the commitment of the governments to open data deployment. The analysis is divided into four indicators: open data policy, open data portals, open data impact and open data quality.

In this line, the report shows that countries that are less advanced in open data policies, tend to engage in the upgrading of their portals, which become the main gateway to open data in the country. More advanced countries as to open data deployment focus, in contrast, on improving the quality of their data publication[23].

Figure 8 Open data (% of the maximum open data score), 2019



Source: European Data Portal.

Figure 3 DESI Report Digital Public Services

4.2 UNCTAD (United Nations Conference on Trade and Development)

The United Nations Conference on Trade and Development (UNCTAD) database provides statistics on international trade, investment and development, with the purpose to provide evidence-based insights for policies and recommendations that would foster economic and social development worldwide. In terms of the creative economy, UNCTAD's data monitors the global trade of creative goods and services, and primarily focuses on the monetary value created through imports and exports. The majority of the time series data dates back to 2002/2003 and the latest data was published in 2015. The Creative Economy Outlook report (2018) summarises this data and provides an overview of global trends in the creative economy in 2002-2015, as well as presenting a detailed analysis of individual country profiles.

[23] See DESI Report: Digital Public Services, pg.8

UNCTAD uses globally-adopted standards and methodologies for statistical analysis and focuses solely on economic impact. This ensures that data can be collected at a global level and easily compared. At the same time, the rigid standards mean that the monitoring of trends in the creative economy is performed at a very high level, therefore failing to provide enough detail to grasp the full impact of the creative sector and to faithfully represent the more recent trends in the creative sector.

For the classification of creative goods, UNCTAD uses the Harmonized Commodity Description and Coding Systems (HS) which is primarily used for the purposes of tariffs, taxation and trade policies. It provides a very extensive and detailed list of creative goods under seven broad categories: Art and Craft, Audiovisual, Design, New Media, Performing Arts, Publishing, and Visual Arts. These categories primarily include physical objects such as printed books or video game consoles, or creative content captured on physical media such as CDs or tapes. It is important to note that no digital products are part of the HS classification. For the presentation of data on the creative services, the Extended Balance of Payments Services Classification (EBOPS) is used. This includes online video/audio offerings, news agency services, online games, software, heritage services, as well as intellectual property fees. However, on the UNCTAD database, this data is bundled up under generic categories, such as “Audiovisual and related services” or “Other other personal, cultural and recreational services”, which do not give meaningful insights into the contribution of different types of services and business models to the economy.

UNCTAD’s adopted categorisation of creative goods is much more extensive than that of creative services. This does not accurately reflect the situation in the creative sector, which is increasingly driven by services, platforms and user-generated content. The Creative Economy Outlook report from 2018 acknowledges these shortcomings and mentions that a new methodology is being developed to provide a more granular and comprehensive approach. Additionally, the UNCTAD database suffers from an evident lack of data gathered from the heritage sector. Many countries have not provided any data from this sector, which consequently marginalises its importance and contribution to the economy. As highlighted by UNESCO’s report The Globalisation of Cultural Trade: a Shift in Consumption (2016), data on travelling exhibitions and cultural tourism could be collected to fill this gap.

4.3 State of the Commons

Creative Commons is the leading organization supporting the global movement for sharing and collaboration. CC create, maintain, and promote the Creative Commons licenses — free, international, copyright licenses that are the standard for enabling sharing and remix of covered content. Creative Commons provides tools and programs that enable sharing on the web (licenses, legal work, and sharing and accessible resources) and a new photo search engine with filters, lists & social sharing. The last Creative Commons report “State of the Commons” has been conducted in 2017 and is an interesting hub of information and data on cultural contents re-use and licenses.

The report shows data with the following highlights:

- CREATIVE COMMONS LICENSED WORKS: 1,471,401,740 in 2017
- MAJOR PLATFORMS SHARING Creative Commons WORK
    - YouTube: 49 MILLION
    - Wikipedia: 46.7 MILLION
    - Deviant Art: 40 MILLION
    - Wikimedia Commons: 36.9 MILLION
    - europeana: 28.7 MILLION
    - Vimeo: 6.6 MILLION
    - Internet Archive: 3.1 MILLION
    - DOAJ Directory of Open Access Journals: 2.7 MILLION
    - Thingiverse: 2.3 MILLION
- COUNTRIES THAT USED CC SEARCH THE MOST IN 2017: USA, UK, Canada, Spain, Germany, Australia with 3,378,330 sessions and 1,500,000 queries.

4.4 Eurostat

Eurostat produced its last report on Culture statistics in 2019. It includes the most recent data available from Eurostat’s online database, Eurobase. The basis is the methodology of culture statistics elaborated by the ESSnet-Culture, slightly modified in recent years by Eurostat’s working group on culture statistics - as presented in the manual Guide to Eurostat culture statistics — 2018 edition. According to the report, Culture statistics 2019 may be broadly split into two parts: three chapters (on employment, enterprises and international trade) concentrate on the economic dimensions of culture, while the second half of the publication focuses more on cultural participation (from the perspective of individuals).

This resume starts from the seventh chapter, which is about the **Use of ICT for cultural purposes** and one of the most interesting for inDICEs’ aims. Eurostat’s statistics on the use of ICT for cultural purposes are gathered from the annual Community survey on ICT usage in households and by individuals.

Here, the most interesting data:

- *In 2018, some 89% of households in the EU-28 had internet access (regardless of the type of connection), this share had increased by 10 percentage points when compared with 2013; In 2018, the share of the adult population (aged 16 to 74 years) in the EU-28 who used the internet during the three months prior to the survey and had watched streamed television (TV) or videos during this period was 72%. This was considerably higher than the corresponding shares registered for listening to music over the internet (56%) or playing or downloading games over the internet (33%);*
- *Reading online news sites/newspapers/ news magazines In Lithuania, Czechia, Croatia, Estonia and Finland, at least 90% of the adult population (aged 16 to 74 years) who used the internet during the three months prior to the survey in 2017 made use of the internet to read online news sites, newspapers and news magazines. On the other hand, this share was less than two thirds of all internet users in Ireland, Belgium, France and particularly Italy (56%); In 2018, some 56% of EU-28 internet users (aged 16 to 74 years) listened to web radio or music streaming services (downloading excluded). More than 70% of internet users made use of web radio or music streaming services in Finland, Sweden and Greece. By contrast, the share of internet users making use of web radio or music streaming was at its lowest in Belgium (43%) and Latvia (47%);*
- *Across the EU-28, some 33% of internet users (aged 16 to 74 years) participated in this cultural activity during 2018. A relatively high share of internet users in the Netherlands (47%), Denmark (43%) and Belgium (43%) made use of the internet for playing or downloading games (as shown by the darkest shade of blue in Map 7.2), while the lowest proportions were recorded in Austria (21%), Bulgaria (22%) and Poland (23%);*
- *Aside from reading online news sites, newspapers and news magazines, young people (aged 16 to 24 years) in the EU-28 were more likely than average to make use of the internet for a wide range of cultural purposes (see Table 7.1). In 2018, some 90% of the internet users in this age group watched streamed TV or videos (compared with 72% of the whole target population and 54% of internet users aged 55 to 74 years), 86% listened to music online (compared with 56% and 30% respectively), while 58% played or downloaded games (compared with 33% and 20% respectively).*
- *In 2017, this was most notably the case for reading online news sites, newspapers and news magazines: 85% of EU-28 internet users with a tertiary level of educational attainment made use of the internet for this purpose compared with 56% among internet users with at most a lower secondary level of educational attainment. One*

was an exception: a higher share of internet users with at most a lower secondary level of educational attainment (39%) made use of the internet in 2018 for playing or downloading games compared with either of the other educational levels;

- *Men were more likely than women to make use of the internet for cultural purposes;*
- *In 2018, less than one fifth (17%) of EU-28 internet users made an online purchase of films or music, a share that reached 22% for online purchases of books, magazines and newspapers, and 27% for online purchases of tickets for cultural or sporting events (it is not possible to make the distinction between these two types of tickets in the Community survey on ICT usage in households and by individuals).*

In the **Cultural Heritage** section, the first one ever featured in this series, the report provides information on European cultural heritage, with data derived from a range of external sources outside of official statistics that are collected by Eurostat, including UNESCO lists, the European Heritage Label, the European Capitals of Culture, EGMUS and a special Eurobarometer survey on cultural heritage (2017). The second chapter on **Culture-related education**, focuses on two areas that link education with culture on one hand, tertiary students who are studying culture-related fields of education; on the other, the role played by education in facilitating cultural exchange, for example, by learning foreign languages or by promoting the mobility of tertiary education students between EU Member States. The third chapter seeks to provide an overview of developments in **cultural employment** and information on the relative weight of cultural employment against the total number of persons employed. The 4th and 5th chapters are devoted to **Cultural enterprises and trade statistics for cultural goods**, providing information on the value of international exchanges of these goods and show the weight of cultural trade within all EU-28 international trade. The sixth chapter is on **Cultural Participation**. It presents some interesting findings about people's involvement in cultural activities analysed by parsing a broad range of socioeconomic characteristics. The last chapter is devoted to **Household cultural expenditure**. Nearly 3% of household consumption expenditure in the EU was devoted to cultural goods and services

4.5 EU Open Data portal

The European Union Open Data Portal (EU ODP) provides access to an expanding range of data from the European Union (EU) institutions. The EU ODP was set up in 2012, following the European Commission Decision 2011/833/EU on the reuse of its documents.

All EU institutions are invited to make their data publicly available whenever possible so they can be used without copyright limitations. All this data is freely available and can be reused in databases, reports or projects. The site offers a number of datasets in various digital formats, that come from EU institutions and EU countries, and are concerned with geographic, geopolitical and financial data, statistics, election results, legal acts, data on crime, health, the environment, transport and scientific research. The most interesting open dataset for inDICEs is the one explorable on Education, culture and sport.

4.6 Main findings from the survey

What our survey of the most relevant sources shows, is the extreme heterogeneity and fragmentation of the collected data and the criteria for collection, as well as the substantial lack of interoperability across different sources. Culture has long suffered from a relative neglect of statistical data gathering compared to other production sectors. The consequence is that - as of today - making a reasonably granular comparative analysis of the main trends and structural features of cultural and creative sectors in Europe is not viable yet.

Based on this observation, one of the purposes of InDICEs is streamlining data collection and analysis, also by means of innovative techniques and tools, to analyze aspects that are generally neglected or not systematically tackled through traditional data gathering activities and through the use of more standard methodologies. The available sources just surveyed are certainly of help in this regard, but inevitably the nature of the results that can be derived cannot fully overcome the limitations and lack of systematicity of the sources.

5. Data gathered up to M12

The inDICEs project objective n.4 is to establish an Observatory Platform to track policies and trends over the long-term, making the collected data openly available on a dedicated and sustainable platform that is connected to the Europeana Digital Service Infrastructure and will offer a participatory space for dialogue and experience exchange for communities and experts in the research, cultural and creative sectors.

For this purpose, inDICEs has collected a group of datasets from the most important cultural production websites, online newspapers and social networks during the first 12 months.
At the methodological level, the choice of collecting data from the selected platforms is due to the intention of covering, for the reasons explained at the end of the previous section, the widest possible spectrum of the different modes of production and types of cultural content present on the web. From this point of view, Wikipedia reflects the desire to analyze a specific type of content that has characterized Web 2.0: the so-called free digital encyclopedias. With regard to social media, a combination of two different types of media has been selected: visual and textual content. In fact, Youtube and TikTok belong to the first category and are respectively the oldest and the most recent social media platforms that drive the spread of online video. Both are supported by different algorithms and produce different types of content. If Youtube can be considered a generalist platform linked to searchable and stable content, TikTok opens the doors to the ephemeral as it specialises in the dissemination of short videos and in the rapid turnover of popular content. In turn, Twitter and Facebook can be considered two leaders of digital information communication; they are based on two different approaches, a rapid and realtime one for Twitter and an opinionated one for Facebook. In addition to the best-known social platforms, a series of purely artistic contents from the repositories of cultural platforms such as IMDb, deviantart and a few digital theatres were collected: in this way it was possible to investigate a sector of the traditional cultural industry that is now present on the web.

5.1 Data gathered at M12

The aim of this first year of data gathering is to provide inDICEs partners with a first group of datasets in order to:

1. explore the redistribution of contents among the digital sphere;
2. validate the data gathering processes and evaluate how to fill the gaps;
3. give a proof of concept on how targeted users can access and utilize this data;
4. advance a preliminary investigation on users behavior;
5. lay the foundation for future researches.

5.1.1 Dataset 1: Crawled Web Content News & Web Sources

The InDICEs crawled Web content is vital to the subsequent data analyses, predictive analytics and content recommendation work. As an initial step, a repository of cultural heritage-related content was built in the Metadata Repository by consortium partner webLyzard through the set-up, configuration and running of various ’data mirrors’ - each collecting at regular intervals online documents matching a search configuration on pre-defined websites or platforms. While the setup currently comprises German, English, French, Dutch and Spanish general news and web sources, an internal data gathering process was initiated as part of WP1 to further extend and customize the covered web content and provide a personalized data feed with targeted content that is of relevance to the cultural heritage sector and creative industries.

The current dataset of news and web sources comprises 19 million news documents and 2.5 million general Web articles, gathered between the 1st of January 2020 and the 15th of December 2020. Split by language, around 6 million of those documents are in English, 7.5 million in Spanish, and 5.6 million in German. The French and Dutch sub-repositories account for 2.4 million and 700k documents respectively. A planned next step is to further extend the content collection to Italian news and web sources.

5.1.2 Dataset 2: Crawled Web Content - Social Media

Complementing the Web sources outlined above, a process was started to gather social media content from multiple platforms including Twitter, Facebook and YouTube. In this first year the focus on social media has privileged Twitter, where 4.6 million tweets, the majority of which are in English with 3 million gathered short texts were collected and analyzed. German, Dutch and French make up the remaining tweets with 800k, 600k and 200k respectively. InDICEs pursues a hybrid strategy for social media content ingestions, based on a combination of domain-specific keywords and specifying important Twitter accounts to capture, for example OpenGLAM, Europeana, UNESCO and the ESC official accounts.

5.1.3 Wikipedia

We implemented a pipeline for Wikipedia data collection, then processed raw data and made use of Azure Batch for the project. Azure Batch is a cloud service to run in parallel several processes in an automated and very efficient way.

In order to collect the data we developed a script that downloads the dump files from the Wikidump portal. The Italian and Spanish data were stored initially on a dedicated Virtual Machine (VM); then, after the processing, they were manually uploaded to the data lake storage with Azcopy, a package designed to interact with the Azure Storage via Command

Line Interface (CLI). For the English Wikipedia, we modified the script to stream the data directly into the storage.

Once the data were collected, we needed to process them because the raw XML files were not suitable for the analysis task. Most of them, moreover, did not contain the refined graph links we wanted. We built on the methods of a recently published article (Consonni 2019) in order to extract the needed information from the *Dump Dataset*.

We decided to adopt the Azure object storage because it offers unlimited storage, and it is object-based as well as cost-effective (see Deliverable D1.1 for detailed process).

To do so, the project builds upon the Culture 3.0 framework, as mentioned in Deliverable D1.1, that was originally developed in the context of the Open Method of Coordination Table on Cultural and Creative Industries and subsequently formalized by a study commissioned by the European Commission to the European Expert Network on Culture.

Today, we are witnessing the coexistence of three regimes of production of cultural content: the patronage regime in its various forms (including public patronage, that is, financing cultural production through public subsidies) (Culture 1.0), Cultural and Creative Industries (Culture 2.0), and open communities of practice (Culture 3.0). The three regimes have materialised in different historical moments in the West but currently, all of them coexist, but different sectors of production of cultural and creative content tend to refer naturally to different regimes. Culture 1.0 basically regulates all non-industrial sectors where reproducibility of content is either not possible or not meaningful: visual arts, performing arts, museums and heritage (therefore also comprising CH sector). Culture 2.0 typically covers cultural industries (cinema, publishing, radio-television, music, video games) and creative industries (design, fashion, industry of taste, architectural design, communication and advertising, serious gaming, etc.). Culture 3.0, finally, is relevant for the new digital platforms of content production and delivery, including social media.

Whereas both Culture 1.0 and 2.0 are based upon the distinction between producers of contents and audiences (with Culture 2.0 massively enlarging the potential audience pool due to technological reproducibility of content), Culture 3.0 is characterised by the progressive blurring of the roles of content producers and users. Due to cheap access and high usability of the new technologies of digital content production in all media (text, multimedia, music, still and moving image, gaming etc.) even non professionals are quickly enabled, if willing to, to produce content with quasi-professional technical standards in relatively little time and cost. At the same time, social media and digital sharing platforms allow a very simple and potentially ubiquitous distribution of such content, of course keeping into account obvious constraints dictated by the economics of attention, by the laws of the experience economy, and by the still crucial role of social salience of content as promoted by advertising, filtering by superstars and influencers, and social centrality of media channels in determining its actual exposure and circulation.

5.1.4 Dataset 4: IMDb

IMDb is a popular platform for movies, TV and celebrity contents, designed to help users explore the world of movies and shows. It was launched online in 1990 and has been a subsidiary of Amazon since 1998.
We prepared an automated script that downloads weekly snapshots of the information in the IMDB website through their API. Due to the successful experience with Azure with the Wikipedia datasets, we decided to store the data there as well. At present, we have 75GB of data, comprising almost 100.000.000 titles, with geographical information, casts and crews, ratings, etc.

5.1.5 Dataset 5: TikTok

TikTok is an online social platform released in 2016 that in just a few years has become one of the most used worldwide, ranking first among the most downloaded free apps for long periods of time.
We downloaded the information corresponding to 25.000.000 TikTok videos (approx. 14GB). For each video we have information at the level of both the user and the video content, such as the video creation time-stamp, the text accompanying the video, number of comments and likes, the creator username, etc

5.1.6 Dataset 6: DeviantArt

DeviantArt is an online social community, founded in 2000. This platform was created to exhibit, promote, and share pieces of art, from literature, painting and sculpture to digital art, pixel art, films, and anime. Nowadays DeviantArt counts over 50 million registered members, known as deviants, and attracts over 45 million unique visitors per month.

We prepared an automated script that, every day, downloads the information regarding the most popular artworks proposed by DevianArt's API, according to their own criteria, and the most popular artworks of fixed categories, chosen by us. We currently have 780GB of data.

5.1.7 Dataset 7: Alltheater.com

Alltheater.com is a subscription-based streaming service which offers online streaming of a library of theater plays. We met with the owners of the platforms in order to explain the project, and they decided to collaborate with us by offering their anonymized data.

5.2 Targeted strategies of dataset analysis

As detailed in Deliverable D1.1, we build a toolbox of inDICEs techniques that allows us to extract different types of quantitative information from texts and provides the basis for sophisticated qualitative assessments to address specific research questions.

The targeted strategies are aimed at investigating data on behavioral responses and change by collecting data from online social media and carrying out analyses of their content and of the users generating it (Wikipedia case study and other datasets).

The targeted strategies for this first year are the following:

- Complex networks
- Mechanistic models of Socio-Physics
- Statistical Modelling
- Web Analytics

6. Preliminary analysis on users' behavioral responses and change

According to one of the most important goals of inDICEs data gathering, in this first year inDICEs WP1 partners started collecting and analysing data on behavioral responses and change, harvesting data from online social media and carrying out semantic analysis of their content.

The scope of this preliminary investigation is to provide possible indicators for feeding the Visual Analytics Dashboard with additional statistical observations and to better understand user behavior on different platforms, investigating in particular behaviors in communities of 3.0 platforms of knowledge, to carry forward research and scientific publications on digital cultural productions and participation. This information will be useful for the inDICEs project to fully understand user behavior of web 2.0 communities, and to design evidence-based socially and economically impactful policies about digitisation of cultural heritage in the future context of the Digital Single Market.

It is thereby possible to track the evolution of the mood and dispositional orientations of users as a consequence of being exposed to certain content. This innovative methodology has been developed by FBK and sets a new standard for the behavioral change dimension of impact evaluation. It is currently employed in a pilot experiment in the cultural field, specifically in the behavioral change component of the impact evaluation conducted on the activities of Matera 2019 European Capital of Culture.

Wikipedia has been chosen as the first case study of special interest and is taken as a benchmark for the deployment of future strategies and approaches.

6.1 The case study: Wikipedia Dataset analysis

Wikipedia is a native Culture 3.0 platform which is of special interest for our analysis, as it represents the most ambitious and articulate example of a decentralized collaborative knowledge platform, as well as a radical alternative with respect to more traditional forms typically developed under the patronage regime, such as conventional encyclopedias. inDICEs can therefore investigate to what extent Wikipedia actually represents an example of a highly collaborative platform that invites the active participation of a large number of users for the prosocial creation of a knowledge commons.

inDICEs implemented a pipeline for Wikipedia data collection; then processed raw data and made use of Azure Batch for the project.

There are many Wikipedia datasets, which are periodically loaded on the Wikidump portal (see https://dumps.wikimedia.org/) and which offer different types of data content. inDICEs focused on the current snapshot of Wikipedia, that is, all the articles and all the links between them. inDICEs was also interested in the complete list of *revisions* (also called *edits*) made by users. This is called the *Dump Dataset,* consisting of a variable number of compressed (gzip or 7z) XML files, produced by Wikipedia at the beginning of every month and made available on their portal or through hosted mirrors. There is a *Dump Dataset* for every Wikipedia language project. For the purpose of the project, inDICEs collected three *Dump Datasets*: Italian, Spanish and English. The size of the datasets are shown in the next table:

Dump dataset	**Size Compressed (GB)**	**Estimated Size Exploded (GB)**
Enwiki	172	17200
Eswiki	20	2000
Itwiki	18	1800

In order to collect the data, inDICEs developed a script that downloads the dump files from the Wikidump portal. The Italian and Spanish data were stored initially on a dedicated Virtual Machine (VM), then, after the processing, they were manually uploaded to the data lake storage with Azcopy, a package designed to interact with the Azure Storage via Command Line Interface (CLI). For the English version of Wikipedia, inDICEs modified the script to stream the data directly into the storage.

6.2 Data extraction and redefinition

Once data was collected, inDICEs needed to process it because the raw XML files were not suitable for the analysis task. Most of them, moreover, did not contain the refined graph links inDICEs needed.
inDICEs built on the methods of a recently published article (Consonni 2019) in order to extract the needed information from the *Dump Dataset*.

The process of extracting and refining of the data was the following:

1. Three intermediate datasets were extracted from the raw *Dump Dataset.*
    a) *Revisionlist*: *list of all the revision made by users*
    b) *Raw Wikilinks*: *list of all links between pages*
    c) Redirects: list of pages that redirect to others (e.g: *NY* redirects to *New York)*

2. Snapshot phase:
    a) The *Revisionlist* dataset is snapshotted for every year and produces the *Snapshots* dataset.
    b) The *Raw Wikilinks* is also snapshotted and produces the *Raw Wikilinks Snapshots* dataset.
    c) The *Redirects* are resolved (in a sort of deduplication of the links)
3. The last phase is the extraction of the *Wikilinks Graph* from the *Raw Wikilinks Snapshots* and the *Resolved Redirects*.

Owing to the large size of the datasets, inDICEs decided to adopt the Azure object storage because it offers unlimited storage; it is object based and cost-effective.

Italian and Spanish *dumps* were effectively processed in a dedicated VM. For English wiki, though, inDICEs estimated a total processing time of 4400 hours distributed on 7 parallel processes, the workload limit for the VM used, that is, about 26 days of work. To speed up the process, and save computation money, inDICEs used an Azure Batch, a massively scalable computation platform that allows for large-scale parallel batch workloads to be run in the cloud. The Batch platform allowed us to distribute the processing workload over multiple VMs.

After distributing the processing into independent single input/output operation tasks, inDICEs defined and configured the execution environment for the batch script by installing the dependencies on the VMs, retrieving the processing libraries from the Github repository, downloading the files to be processed on the VMs (one file per machine at a time) from the Azure storage and instrumenting the scripts to upload the results to the cloud.

In addition to the list of edits made by users (corresponding to the *Revisionlist* dataset) and the graph of the links between the pages (*Wikilinks Graph*), inDICEs decided to store the other intermediate datasets as well. The following table shows the final dimensions of the various datasets produced during the processing for the three languages.

Dataset	**Enwiki**	**Eswiki**	**Itwiki**
Raw Wikilinks	746	111	92
Raw Wikilinks Snapshots	36	8	9
Redirects	0	0	0
Resolved Redirects	4	1	1
Revisionlist	62	9	8
Snapshots	42	8	6
Wikilinks Graph	21	4	4
Total	911	141	120

Tab: Final output datasets, with their dimension (compressed GB)

As a final note, the inDICEs WP1 working group wants to remark that the data processing was not fully automated and required some intervention and monitoring, in two different ways. On the one hand inDICEs used the statistics files that the scripts produced, containing information regarding the number of pages parsed, the number of revisions analyzed, the time taken etc.; on the other hand, the Batch Explorer client was widely used. With this tool it was possible to manually scale the size of the VM pool to avoid having a series of machines in idling and thus limit the expenses. It was also possible to monitor the general progress of the tasks, identify any problems and manually intervene to resolve them.

7. Plan for the next period

In the following section, the plans for the next 6 months [M15-M21] are summarized. The plans regard firstly the WP1 work, which will be integrated and eventually re-shaped mainly with the collaboration of the WP4 regarding the Open Observatory tool integration, the WP5 for the Open Call dissemination, and all the members of the consortium more generally.

The plans are aimed at the full compliance with the D1.2 Data register [M18] objectives, which regard the definitive design of the structure and characteristics of data gathered, and of their sources, and a discussion of their usefulness and limitations, within an organizational scheme that allows their effective accessibility into an Observatory Platform.

7.1 Repository

During this first year of data gathering and analysis, WP1 has deployed all the forces and professionalism present within the consortium of inDICEs partners to pursue the main objectives, always taking into account the possible limitations emerging in the various steps during the progress of the project.

In this perspective, it proved necessary to introduce within the Open Observatory an additional tool, identified during WP1 internal and collective meetings as a Repository.

Within the inDICEs participatory platform, in conjunction with the Visual Analytical Dashboard, a data repository will be realized as an appropriate, subject-specific location where inDICEs users can directly upload and access their data.

It will be a place that:

- collects datasets and indicators;
- makes them available to use;
- and organizes them according to the inDICEs different types of targeted users' (personas') needs;
- will provide specific features that will allow users to be facilitated in searching per subject or research domain;
- will face issues concerning data re-use and access; file format and data structure; and the types of metadata that can be used.

It will be developed taking into account the co-created criteria aimed at defining the most desirable structure for the Open Observatory, emerged during the Hypothesis Assemblies; in particular, according to the following criteria:

- C1 Serves, aggregates and manages collected open data and methodological tools;
- C3 Provides a transparent infrastructure to share legal and technical documentation and training resources;

- C5 Ensures a responsible use of user validation, filtering tools, content moderation, and AI techniques to synthesize large volumes of communication and data;
- C7 Creates connections and networks that enable use of work and processes to support resources, tools, strategies, and policies for more effective and cohesive digital transition by CHIs;
- C8 Empowers CHIs to have a stronger voice regardless of size or geography and have a more interdisciplinary approach towards their own data and resource creation.

inDICEs data and indicators repository will be organized in a directory with:

1. sub-folder with data explanation, reports and descriptive boxplot for data visualization
2. sub-folder with open datasets .csv format
3. sub-folder with available useful indicators organized in relation with targeted users' potential needs, as illustrated in Deliverable D4.1 p.24
4. the available integrated form of the Self-Assessment Tool on CHI digitization

In the upcoming period, it will be discussed the users' uploading technical process and how to engage users in creating a community of interest aimed at sharing and working on open-access datasets.

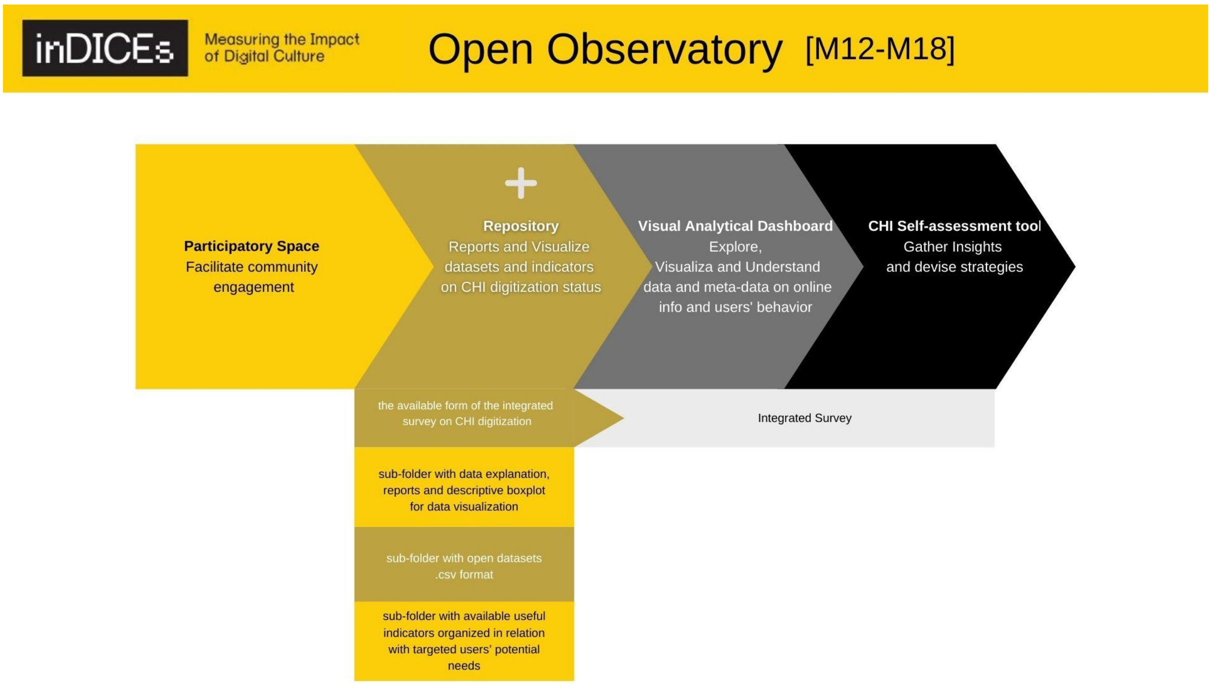

inDICEs
Measuring the Impact of Digital Culture
Open Observatory [M12-M18]
Participatory Space
Facilitate community engagement
+
Repository
Reports and Visualize datasets and indicators on CHI digitization status
Visual Analytical Dashboard
Explore, Visualiza and Understand data and meta-data on online info and users' behavior
CHI Self-assessment tool
Gather Insights and devise strategies
the available form of the integrated survey on CHI digitization
Integrated Survey
sub-folder with data explanation, reports and descriptive boxplot for data visualization
sub-folder with open datasets .csv format
sub-folder with available useful indicators organized in relation with targeted users' potential needs

Figure 4: Repository integration flowchart

7.2 Data acquisition on CHI digitization

In order to support a long-term sustainability of data gathering processes and to align with the already existing reports, in the next six months [M15-M21] data on digitisation of cultural heritage advanced by CHIs will be gathered according to the first list of indicators that have been already identified and to a new list of indicators that will be defined with the help of a group of inDICEs targeted users, such as researchers, CHI practitioners and policy makers, who will be consulted during the inDICEs 2nd Consultation Workshop (20-04-2021).

In addition to content collected and made accessible via the Visual Analytics Dashboard, WP1 and WP3 will work together for integrating the Self-Assessment Tool with additional indicators that will produce the specific data needed to align inDICEs work with the already existing reports on CHI digitization. The Self-Assessment Tool integration addressed to the CHI sector will be supplemented with:

- indicators drawn from data analysis on web 2.0 platforms' users behavior
- indicators/questions drawn from data analysis on WP2 survey on intellectual property rights and CHIs e.g. questions on IPRs, IPR status of CHI's collections, licenses used, etc.
- indicators constructed in order to fill the most relevant missing information on realms e.g. participation, new professional figures on cultural digitization, new web channels and platforms, new strategies of cultural production etc.
- targeted indicators extracted by the most important statistical reports

The set of data gathered will contribute to both customize the Visual Analytics Dashboard and provide content for the Repository.

7.3 Configuration Template

To gather data of relevance for the CHI sector and customize the content feed, the keywords and topics for the Visual Analytics Dashboard will be reevaluated and updated throughout the project. The keywords classification carried out in this first year of work provides a list of the most relevant terms with a twofold scope: they act as a filter for the web crawling that generates the content accessible through the Visual Analytics Dashboard, and help users conduct their research. The current list includes terms that need to be disambiguated and refined in the next period in order to better understand and re-discuss the process of filtering. Once the first scraping of the digital traces inherent to the different models of cultural production has been completed, it will be necessary to carry out a preliminary and exploratory analysis regarding their real relevance to the different phenomena under investigation. In fact, the inductive-deductive process typical of digital methods involves a continuous reassessment of pre-established beliefs regarding how such phenomena actually emerge in online debates. From this point of view we foresee that WP1 will be able to carry out in the next months a process of reduction and enrichment of data collection techniques, in order to allow a better characterization of the same data.

7.4 Open Call for Open Sources

In order achieve some of the most important inDICEs project goals, which are effectively engaging, not only inDICEs partners, but also communities in participative research activities, and developing strategies of democratization of cultural production and access, according to the inDICEs theoretical frameworks which build upon the 3.0 regime of cultural creation that is based on bottom-up and collective co-creation of contents, inDICEs is about to launch an Open Call for Open Sources.

The aim is to populate the inDICEs Visual Analytics Dashboard, a tool for exploring online content and tracking recent trends by putting data in context along multiple different metadata dimensions (extracting keywords, sentiment, relations and geolocations), with a growing and constantly updated amount of on-line open sources of cultural production and reproduction. Doing so, the inDICEs Visual Analytics Dashboard and the Observatory Platform will be nourished not only by top-down sources selected by a group of experts, but also by a bottom-up process of co-creation. The Open Call will address not only the inDICEs partners and community, but also a huge cross-national group of already targeted users (personas), who can contribute to inform Observatory administrators on the most interesting and valuable open sources about on-line cultural contents and off-line re-utilized sources and case-studies.

This can be useful to give the inDICEs Observatory a wider look upon the most interesting and local cultural sources, without missing the chance to detect and involve even the small heritage communities or minorities in cultural production and reproduction sources.

Users will be asked, via a simple survey developed on the inDICEs Participatory Platform, to add any web sites, Facebook pages and Twitter handles whose content they would like to have monitored and analyzed.

The conditions for contribution will be:

- web sites can be of the own institution, partnering institutions, or simply web resources frequently consulted for daily research or considered relevant for inDICEs
- text-based web sites are preferable, e.g. news pages, blogs etc.
- web sources need to be open-access

The users will be asked

- to select their favorite top 1 to 5 online sources of cultural contents
- to categorise them through a drop-down menu in the following pre-structured list of fields of CCS:

❏ Contemporary art
❏ Modern art
❏ Ancient art
❏ Archeology
❏ Archives (national, records office, audio-visual/broadcasting archive)
❏ Theatre (any kind)
❏ Performing arts
❏ Cultural institutions
❏ Museums (art, archeology or history, natural history or natural science, photography, science or technology, ethnography or anthropology, civil/human rights)
❏ Art market (galleries, auction houses, art fairs)
❏ Architecture and design
❏ Library (national library, higher education library, public library, special, private or other type of library)
❏ institution for monument care
❏ University / research institute
❏ Cinema (Film labels, Short films/self production films, Sub-culture films)
❏ Radio/podcast/music forums
❏ Street culture/subcultures fan bases
❏ Fashion
❏ Design
❏ Participation/cooperation cultural practices
❏ Meme or contemporary re-use of cultural contents
❏ Hacker/hacking and systems of cultural subversion
❏ Public art
❏ Activism and artistic/cultural practices

Legal online sources on CHI:

❏ Intellectual property scientific journals
❏ IT law scientific journals
❏ Law and arts scientific journals
❏ Human rights/international law scientific journals
❏ Intellectual property blogs
❏ IP news from international organizations (e.g. WIPO, EUIPO)
❏ Blogs/news from organizations of art and law

❏ intellectual property rights
❏ copyright law
❏ trademark law
❏ databases rights
❏ designs law
❏ traditional cultural expressions
❏ interactions of arts and law
❏ interactions of new technologies and law
❏ digital policies
❏ fundamental/human rights
❏ international law

The internal process will be the following:

1. Sources will be gathered through the survey
2. WP1 will evaluate the most valuable sources
3. Sources will be added to the Configuration Template
4. The Visual Analytical Dashboard will benefit from an increased pool of sources to extract content and enrich it with additional metadata
5. Users will be able to filter domain-specific documents with personal queries

Management and dissemination strategy:
WP1 will test the correct functioning of the survey by sharing a preview with inDICEs partners and, after the kick-off that will be held during the inDICEs Consultation workshop “Developing Future Researchers”, will integrate results into the project, selecting the most relevant sources added. WP5 will structure the graphics, the communication plan and the dissemination of the project.

ANNEX 1. Museum statistics collected by EGMUS and NEMO

Table 1 concerns the total number of museums. Data are partial. Some countries include only some categories of museums, for example Italy and Greece report only state-owned museums; actually, the notes to the EGMUS table state that Greece reports only archaeological museums.

Table 2 reports data about basic IT technology diffusion in museums: availability of a computer and of a museum website. Only 10 countries replied to these questions, so the totals over the EU27 countries give little information.

Table 3 presents the responses to the NEMO Survey compared to the number of museums, by country.

Table 1 – Museums in EU27 by type of collection, ownership and management

EU27	Year	Number of museums according to type of collection				Ownership				Management			
Country		Total	Art. archaeology and history	Science & technology, ethnology	Others	State-owned	Local/regional-owned	Other public-owned	Private-owned	State-managed	Local/regional-managed	Other public-managed	Private-managed
Austria	2017	549	187	64	298	38	219		292	3	139	60	347
Belgium	2004	162	192	58	56	8	74	6	47	8	53	7	76
Bulgaria	2017	191	140	16	35	25	162	4					

Croatia	2014	284	114	61	109	33	167	83	1	33	167	83	1
Cyprus													
Czech Rep.	2019	286	108	53	325	30	350	47	59	30	350	47	59
Denmark	2017	209				5	17		187				
Estonia	2018	250	58	23	169	72	90		88	65	90	7	88
Finland	2019	327				17	215	7	88	17	215	7	88
France	2017	1 224	805	341	78	70	994	26	134	59	983	55	23
Germany	2018	6 741	1 229		5 512	429	2 596	438	3 020				
Greece	2007	176	176			176				176			
Hungary	2018	732				104	476		152	104	476		152

Ireland	2014	230								15	23	10	48
Italy[24]	2017	472				472				472			
Latvia	2019	153				40	97	10	6	40	97	10	6
Lithuania	2019	107				19	55	23	10	19	55	23	10
Luxembourg	2012	54	20	17	17	6	11	4	33	6	9	3	36
Malta													
Netherlands	2016	694	496	158	40	61			633				
Poland	2017	949	294	135	520	78	612	63	196				
Portugal	2017	680	269	170	241	517			163				
Romania	2018	787	266	211	272	76	585		104				

[24] Since Italy provides only the number of state-owned museums, an estimate of 900 museums in total will be used in what follows.

Slovakia	2017	159	63	45	51	57	84		19	57	84		19
Slovenia	2018	93											
Spain	2018	1 461	678	407	376	169	873	32	387	97	884	55	425
Sweden	2018	370	113	28	229	28	93	34	45	41	128		201
TOTAL EU27		17 340	5 208	1 787	8 328	2 530	7 770	777	5 664	1 242	3 753	367	1 579

Table 2 – IT in museums

EU27			Museums equipped with at least one computer								Museums with a web site		
Country	Year	Total number of museums	Equipped with at least one computer	*Percentage on all museums*	For admin, purposes	For visitor's information purposes	*Percentage on all museums*	Having an electronic inventory	Having Internet access	*Percentage*	Possessing a web-site	Possessing an own web-site	Updating themselves the web-site

Austria	2017	549	399	*73%*	116	246	*45%*	214	173	*32%*	430	361	
Belgium	2004	162	112	*69%*	83	34	*21%*	80	102	*63%*	114	68	
Bulgaria	2017	191	190	*99%*	187	79	*41%*	140	187	*98%*		126	
Croatia	2014	284						181	247	*87%*	227	121	121
Cyprus													
Czech Rep.	2019	286				123	*43%*					447	
Denmark	2017	209											
Estonia	2018	250	180	*72%*	180	50	*20%*		172	*69%*	218		
Finland	2019	327	153	*47%*	153				153	*47%*	153		
France	2017	1 224											

Germany	2018	6 741											
Greece	2007	176							176	*100%*	176	2	2
Hungary	2018	732				219	*30%*	217	441	*60%*	312		
Ireland	2014	230											
Italy	2017	472									472		
Latvia	2019	153	152	*99%*	146	94	*61%*	133	151	*99%*			
Lithuania	2019	107											
Luxembourg	2012	54											
Malta													
Netherlands	2016	694									694	694	

Poland	2017	949									812		
Portugal	2017	680											
Romania	2018	787											
Slovakia	2017	159	35	*22%*					23	*14%*	122		
Slovenia	2018	93	93	*100%*	93			93	93	*100%*	93	93	
Spain	2018	1 461	1 445	*99%*	1 221	571	*39%*	713	833	*57%*	1 347	776	
Sweden	2018	370											
Total EU27		17 340	2 759	*…%*		1 416	*…%*		2 751	*…%*	5 170	2 688	123

Table 3 – Responses to NEMO Survey compared to the number of museums, by country

Country	Answers to NEMO survey	Number of	% of answers on total of	% on total number of museums	Ratio of NEMO percentage to

		museums	NEMO survey		museum number percentage
Austria	124	549	15.8%	3.2%	5.0
Belgium	42	162	5.3%	0.9%	5.7
Bulgaria	3	191	0.4%	1.1%	0.3
Croatia	13	284	1.7%	1.6%	1.0
Cyprus	4		0.5%	0.0%	NA
Czechia	23	286	2.9%	1.6%	1.8
Denmark	19	209	2.4%	1.2%	2.0
Estonia	15	250	1.9%	1.4%	1.3
Finland	38	327	4.8%	1.9%	2.6
France	30	1 224	3.8%	7.1%	0.5
Germany	75	6 741	9.5%	38.9%	0.2

Greece	24	176	3.0%	1.0%	3.0
Hungary	15	732	1.9%	4.2%	0.5
Ireland	5	230	0.6%	1.3%	0.5
Italy	33	472	4.2%	2.7%	1.5
Latvia	32	153	4.1%	0.9%	4.6
Lithuania	33	107	4.2%	0.6%	6.8
Luxembourg	10	54	1.3%	0.3%	4.1
Malta	7		0.9%	0.0%	NA
Netherlands	30	694	3.8%	4.0%	1.0
Poland	16	949	2.0%	5.5%	0.4
Portugal	26	680	3.3%	3.9%	0.8

Romania	28	787	3.6%	4.5%	0.8
Slovakia	2	159	0.3%	0.9%	0.3
Slovenia	16	93	2.0%	0.5%	3.8
Spain	62	1 461	7.9%	8.4%	0.9
Sweden	62	370	7.9%	2.1%	3.7
Total	787	17340			