Technical Challenges and Approaches to build an Open Ecosystem of Heterogeneous Heritage Collections

Ricard de la Vega & Natalia Torres & Albert Martínez; Consorci de Serveis Universitaris de Catalunya (CSUC); Barcelona; Catalonia/Spain

Abstract

Empowering Communities with a Heritage Open Ecosystem (ECHOES)[1] is a project that intends to provide an open-source and modular architecture to gather different digital contents related to European heritage.

When a wide amount of heterogeneous data of collections, disciplines and countries are joined, different kinds of technical challenges must be considered. In this article are detailed these challenges and the approaches that have been used, with more or less success, to solve it.

Introduction

ECHOES provide an open, easy and innovative access to digital cultural assets from different nations and languages. Within a single and integrated platform, users have access to a wide range of cultural heritage items that can be explored according to different criteria. The platform works as a digital ecosystem formed by a wide range of user communities and allows an active participation with the ability to enrich the digital collections it contains.

The project is carried out by the Erfgoed Leiden en Omstreken (ELO), Tresoar, the Department of Culture of the Generalitat de Catalunya and the Diputació de Barcelona. The Consorci de Serveis Universitaris de Catalunya (CSUC) participates in the project as a technological partner. Having an initial length of two years (2016-2018), it has been decided to extend the initial development for at least one more year. All the code developed in the project is open source under a MIT license and it is available from GitHub[2]. An open source community is also created to manage the maintenance and further development of the software.

ECHOES encourage that the content of heterogeneous collections of different institutions or even countries can be linked. For example, the work of the Catalan architect Antoni Gaudí is mostly located in Catalonia, but he also participated in some projects throughout the Spanish geography, such as El capricho in Cantabria, Casa de los botines in León or El Palacio Episcopal in Astorga. So, his work is part of different collections managed by the different departments of culture of each region and need to access separately. If we can transform all these collections into the same standard and load them together somewhere, we could find all the works of the artist, searching by different architectural styles (neogothic, modernist) or type of elements used (mitral arches, gusset, etc.). Adding new data from other collections located around Europe we could find new interesting information like his collaborators works and even works influenced by his style, for example the Het Schip from Amsterdam by Michel Klerk.

The theory is sound but there exist many challenges to tackle that are detailed in this article. These challenges, and the approaches that have been chosen are the first part of the paper. The second part of this paper covers the technical architecture, the development is explained, some lessons learned in the first two years of the project are described, and finally, the current and future development is detailed.

Challenges and Approaches

The objective of the project is the cooperation across heritage disciplines, institutions and borders to the European challenge of existing cultural heritage fragmentation (silos). This is a project of interoperability between different data collections. Integrating data is not just about putting them together in a repository, but also to facilitate their access so it can be properly exploited by the public.

ECHOES also aim to provide an easy way for smaller cultural heritage owners to transform their collection into linked data and share their collection data with Europeana.

After reviewing similar projects on big data, the choice was made that all the records inserted into the system should have the same structure and format. This was chosen to simplify the re-use of the data by the public. Transformation of the data into the same structure and format makes it so that the quality of outputs is determined by the quality of the input. If garbage comes in, then garbage comes out. The cleaning and homogenization of large different datasets that are being integrated is an ongoing research topic[4], meaning that this is not an easy task and is requiring a lot of efforts. Linked data suffers from quality problems such as inconsistency, inaccuracy, out-of-datedness and incompleteness. It is therefore important to assess the quality of the datasets that are used in linked data applications before using them[5].

There are two ways to ensure the coherence and consistency of the data: a priori or posteriori. A priori data integration means that new data should be cleaned and transformed to some standard before it is added to the final database. A posteriori integration means that data is added to the system as it is, being cleaned and transformed to some standard in real time at the time it is used. Due to the complexity, and specially to the volume of data to be integrated on the ECHOES project it has been decided to process data at the time is being inserted in the dataset (a priori approach).

Below are the details of eight of the challenges the project identified. The first four refer to the input of data. The next two refer to the output. And finally, there are two challenges that are not easy to classify as input or output. After explaining the details of the challenge, the approach of the ECHOES project to solving this challenge is detailed.

1. Different input collections can have **different input metadata schemas**, such us Dublin Core, A2A, EAD, etc. It was necessary to have a one standard that was the basis for mapping for the rest. The standard that was chosen was the Europeana Data Model (EDM[6]), the most used standard for heritage conten purposes. This choice also facilitates the export of ECHOES hubs contents to the Europeana portal.

   In a first approach the inputs were mapped directly to EDM in the data source module and then passed directly to the data lake (the data repository). Initial mappings were closely related to the first test collections, and this caused chaos due to the quality of the source data. It was necessary to define which data formats are accepted and the mapping, metadata to metadata, from these formats to EDM to assure that only the data that are consistent can be published into the data lake.

   The transformation to EDM of the following formats have now been developed or validated: Dublin Core (DC), A2A, EAD, custom metadata schema from Memorix, custom Catalan metadada schema, Topx and CARARE. This approach is easily extensible, if someone wants a format that is not on this list, creating their own EDM mapping is possible and can contribute to the community.

   One subpoint within this challenge is when a standard has an official conversion, such as CARARE, but is not yet completed. In this case, if a collection with this schema wants to be incorporated to the data lake, there are two options: to wait for the CARARE working group to complete the official conversion, or to self-complete the mapping, running the risk that of not coinciding with the official version. In that case, it must have been adapted again, to adapt to the standard. Both options are not desirable.

2. With the transformation of the first heterogeneous collections of multiple collection owners within the project, without homogenization nor validation processes. It came apparent that that a **poor data quality** limits the exploitation of the data. For example, one unique field which contains references to locations has values at a different geolocation level, there are references to a municipally such as Bussum, a city such as Chicago or a country such as China. The same happens with dates, we have also found problems since some records refer to a specific day and time, another to a whole year, and others referencing to temporal time spans such as centuries. Another quality problem is the misspelling, as we found examples of "Leide4n", "Leideb", "Leidedn", that can be assumed as "Leiden".

   Two modules have been developed to improve the quality of the data. One is focused on data profiling, analyzing the inputs with information about the data types, the number of instances, blank cells, etc. And the other is focused on validating the input data. This second module, called "Quality assurance", reviews each item and based on a defined rule decides if this item can be loaded into the data lake. This module consists of three primary functions: the first, that reviews the EDM schema (tags, mandatory fields…), the second is a semantic validation based on the EDM Schematron and the last part is a validation of the content, based on a pre-defined and configurable list of rules. An input collection can only be loaded into the data lake when the input collection passes the three validations.

   In any case, an input collection that succeeds all the validations is not a complete guarantee that the data quality is of an exactable level. We found examples of records that uses the same metadata field called "coverage", for either information about places or information about dates. There is no common rule that can be added to the validation to counter these human errors in the input collection.

3. Same data from different collections can be stored once and updated. An important task in working with data from different sources is the **data deduplication**. As shown in Figure 1, different collections can include the same object, and store different interesting metadata related to it. Deduplication process needs to detect this and create a unique item that includes all metadata. This process can be performed easily if the objects have identifiers, but it is not usual in heterogeneous collections of different sources. In this case, different similarities and distance metrics algorithm can be used to find duplicates, for exemple, Levenshtein, Jaro Winkler or others implemented with Duke[7] tool. It is possible to adjust the degree of similarity, and to reduce the risk of false positives, comparisons can be made through multiple metadata.

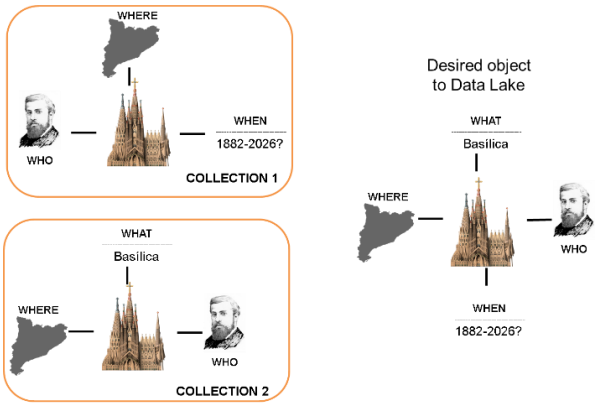

***Figure 1**. Inputs deduplication challenge*

4. One of the most important challenges of the project is the content enrichment. There are two types of enrichments, those suggested by users, and those that can be done automatically. The **automatic enrichments** allow the use of information systems, usually in the form of linked open data (LOD), that can be a useful complement to the original data. For example, possible automatic enrichment for places are TGN, GeoNames or Pleiades, for Agents, VIAF or Wikidata.

To integrate automatic enrichments, it is needed to decide which metadata fields are candidates for the enrichments. After that, it is necessary to decide if the new metadata can be incorporated as a pre or post process, to choose the data sources and, finally, how to do this, reusing existing fields or creating new metadata.

For this challenge a final approximation has not yet been decided. It is a topic to work on during 2019. Some tests have been performed with places. For example, the A2A collections do not have geolocation coordinates, and there have been added based on texts such as "Leiden" from GeoNames. The coordinates are necessary in order to be able to locate data onmaps.
Another example has been performed adding information from DBpedia based on the location, too. It is pertinent to remember that the enrichments depend on the data quality, a challenge that has been previously detailed.

Once data inserted in the data lake has the necessary quality, the challenges for its exploitation will be detailed below. Exploiting integrated systems such as ECHOES in a visual way is a very difficult task, since the heterogeneous nature of the data as well as its volume makes it very hard to represent it. Despite being technically complicated, searching through all the data lake may be interesting in order to find related information from several collections that without this aggregation before it was not possible to find. On the other hand, a lot (perhaps most) of the data it does not make sense to consult it for in an aggregated way.

Instead of focusing on a single option for the exploitation of the data, several alternatives have been studied and tested, which are detailed below. Additionally, a key aspect of the project has always been to offer open access to data so that, via API or publishing as linked open data, the information can be consulted and exploited by other systems.

5. In an attempt to visualize data in the form of a graph, from four random collections on old registers, two from ELO and two from Tresoar, with a total of approximately 167 thousand items. It became clear how challenging **too much data** can be. One way to facilitate the navigation through this content of linked resources is to provide an explorable representation of the underlying graph of data. These data produced a graph of approximately 455 thousand nodes, generated with Gephy. In theory, it was navigable but in practice, as can be seen in figure 2, it was impossible, and it did not provide any type of value. And even worse, to generate it, it was necessary to use high performance computing (HPC) facilities.

   For this challenge, the divide and conquer strategy can be used. The main point is to pick a particular datapoint as a focus for analysis and the system then computes and displays an "optimal" relevant context given to the users current interests[7]. It all starts with a query that generates the first node, with a set of connections to other relevant nodes that allow users to interact with this initial subgraph and expand it in any direction. This way allows to navigate thought any kind of RDF-based system.

   The project is still far from achieving this holistic navigation approach for all data. Not only it is difficult to decide on the entry point, but also how to calculate the relevance of the nodes.

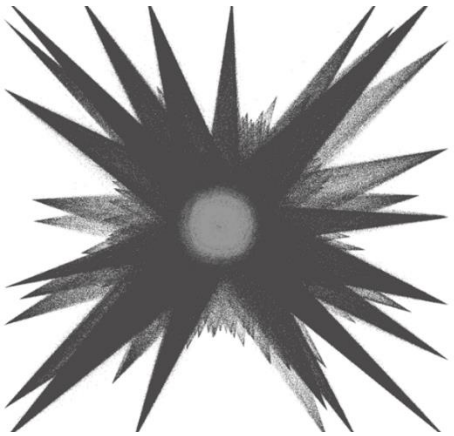

***Figure 2**. Data visualization of a 455.000 nodes graph*

   One alternative for tackling the problem of visualizing integrated data is to define domain-specific visualization[4]. This is, instead of targeting the exploration of the whole database, very specific solutions can be built to solve specific tasks in specific domains, following this approach, visualization tools have been developed for the different types of information to be displayed.

   Cultural objects are shown as a graph that relates places, people, dates and concepts (see figure 5); clicking on each of them the user can browse through data and explore the content. Place searches can be done using a text query box or selecting an area on a map, results are shown also on a map using their geolocation (see figure 7). Dates and periods are represented using timelines where the user can optionally add temporary periods to help to contextualize the date result, for example, the industrial revolution in the Netherlands or in Catalonia are in different year periods (see figure 9); another approach offered for dates is the use of heatmaps to visualize dates because it helps users to focus on dates when various events occur, for example, a peak of deaths could help the user to relate them to an epidemic (see figure 8). Graphs are used to show relationships between people (see figure 6).

6. All the data lake contents are accessible in RDF format through a linked open data endpoint. A user-friendly interface called YASGUI to access this endpoint is integrated. One factor that currently limits the success of linked data repositories is the requirement of knowing the querying language SPARQL and the exact structure of the used database. For all users (most) who are not accustomed to doing queries with the SPARQL language there exist the possibility to develop a visual query system that assist in the generation of SPARQL queries effortlessly. The tool Visual SPARQL Builder

was tested with the ECHOES data (see figure 4). The system allows to drag any of the elements from the database to an infinite canvas, showing the metadata of each element in boxes to specify details of the query. Then relations between boxes can be created in order to combine the information from different elements.

The project challenges are not only related to inputs and outputs, but there are two more challenges difficult to classify.

7. The software developed in the project must be useful for **different scope** institutions, from small to great national or international hubs. The technology developed is scalable so that it covers many different scenarios, from small collections to instances including collections from many institutions. It is possible that small instances do not use all the developed software modules. On the other hand, it is necessary to size the large instances well and perform performance tests.

8. It has already been said that one of the objectives of the project is the **user enrichment** of the contents. This challenge has all the difficulties mentioned above for automatic enrichment. Although some initiatives such as Zooniverse have been analyzed, priority has been given to the provision of mechanisms for the processing of input data and their subsequent exploitation mechanisms, as a basis for doing this functionality afterwards. It is planned to start this issue the second semester of 2019.

Technical architecture

The architecture of ECHOES is formed by a modular approach, with four main pieces (see figure 3) the mapping and transformation data sources module, the data lake, the data retrieval and visualization module and the enrichments. The core module is the data lake, where the data from heterogeneous collections (inputs) are introduced after a process of cleaning and transformation. Once the "normalized" data is available in the data lake, this can be searched and accessed in different ways (outputs). Finally, the data in the data lake can be enriched, automatically or with the user collaboration.

The data sources module is formed by four pieces to transform the input collections to data with enough quality to be introduced into the data lake.

1. The **Analyze** optional submodule enables data profiling. It generates reports that summarizes the contents of the data source.
2. The **Transformation** optional submodule is the tool for data standardization, to prepare data sources to EDM standard. Nowadays, there are the following transformed metadata schemas: Dublin Core, A2A, EAD, Custom from Memorix, Custom Catalan metadata schema, Topx and CARARE. (see figure 10).

3. The **Quality Assurance** submodule, as its name implies, reviews each item. And based on defined rules decides whether it can be loaded into the data lake. There are three kinds of data validations: the syntactic EDM schema, the semantic EDM schema with Schematron and a content validation.
4. The **Publish** submodule oversees the data deduplication and loads datasets into the data lake.

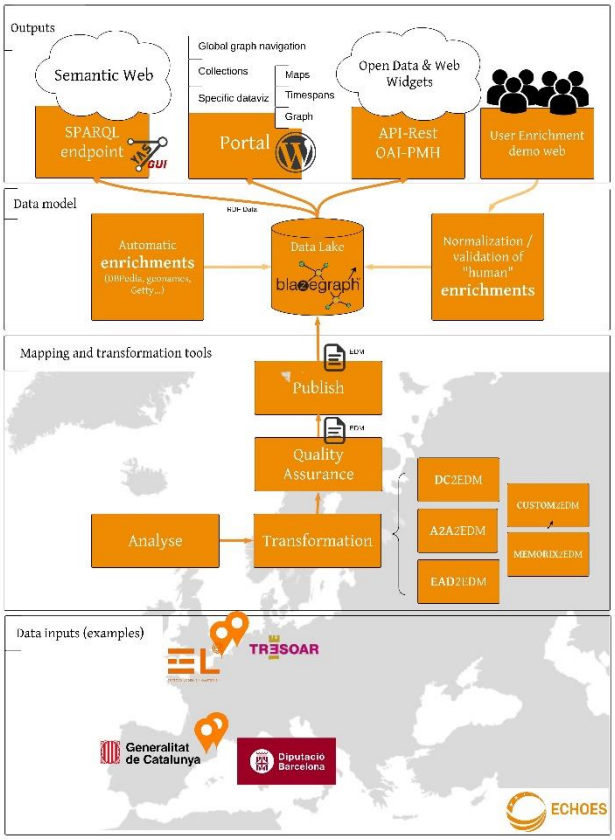

Figure 3. ECHOES technical architecture

The data lake module contains a big amount of data from different sources in EDM. It is built by a graph open source database called **Blazegraph**.

The data retrieval and visualization module is composed of modules to exploit the data in the data lake in different ways.

1. A SPARQL endpoint called **YASGUI** that opens the collections and link them to the world as linked open data (LOD).
2. A web **portal** builder with a modular and extensible architecture. At present it is built with the WordPress CMS with some additional custom plugins. Some data visualizations detailed before are implemented on this, such as a map, timespan, heatmap or graph visualizations.
3. Mechanisms to export the data lake contents via a REST API or with the OAI-PMH protocol.

Only some test was done with the last part of the architecture, the **enrichment module**, that includes automatic and manual enrichments.

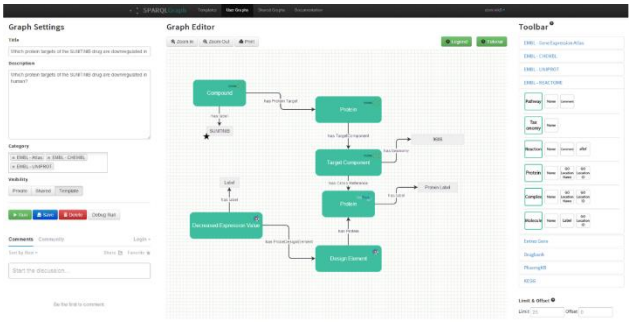

***Figure 4**. Visual SPARQL Builder*

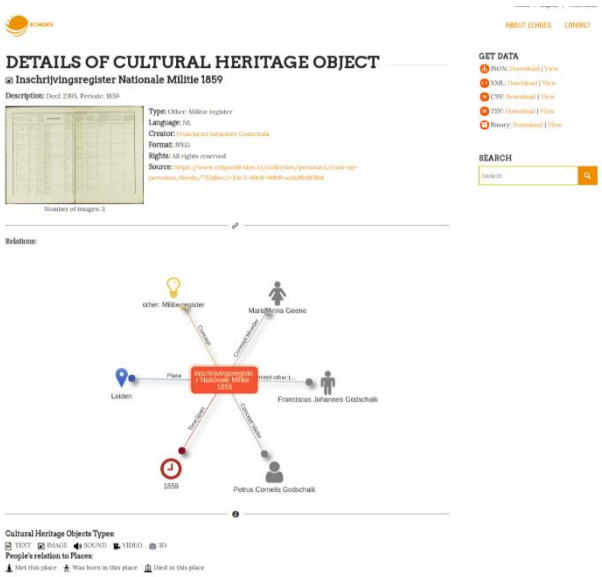

***Figure 5.** ECHOES graph data visualization of concepts*

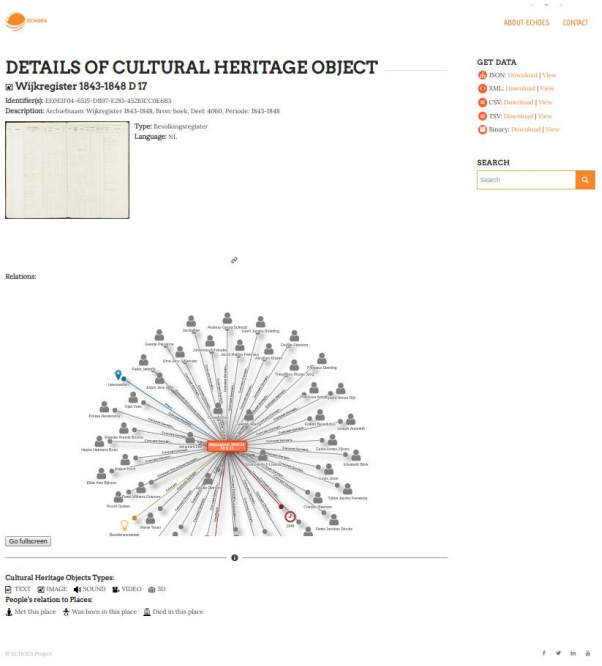

***Figure 6.** ECHOES graph data visualization of people relations*

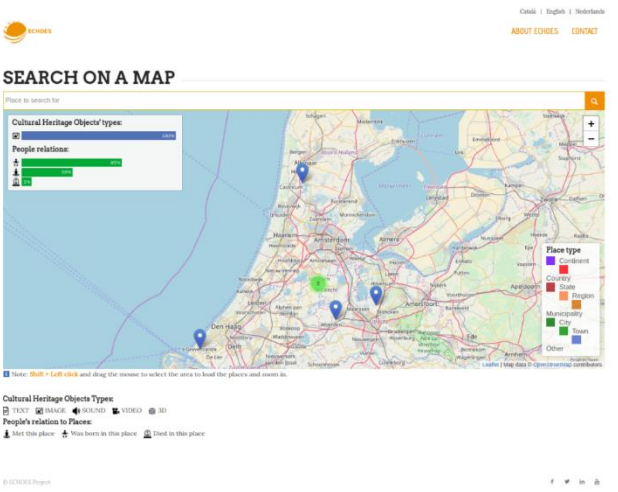

***Figure 7.** ECHOES map data visualization*

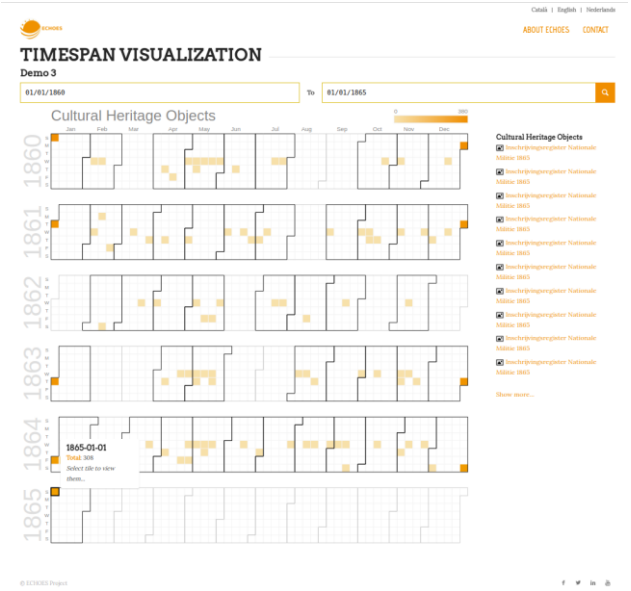

***Figure 8.** ECHOES heatmap data visualization*

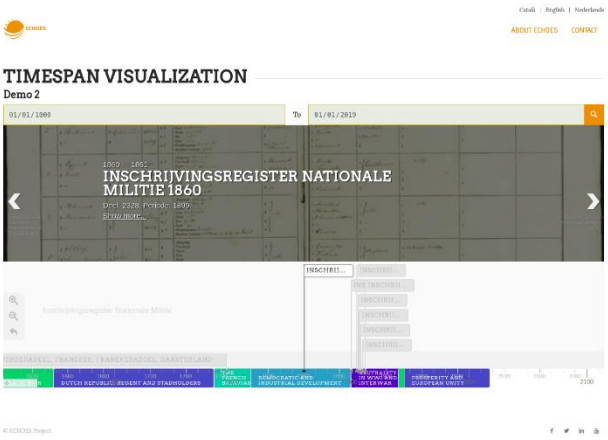

***Figure 9.** ECHOES timespan data visualization*

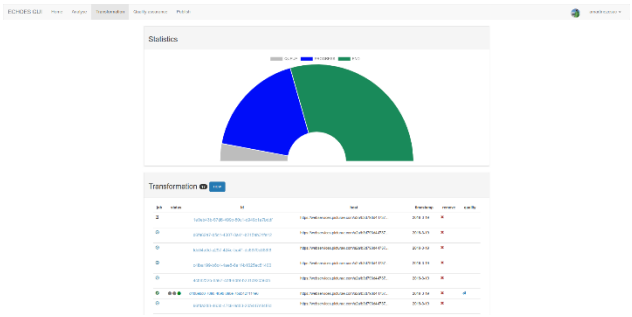

Figure 10. *ECHOES GUI transformation module*

Lessons learned (for now)

There are methodologies that have been useful during the first years of the project, likewise, there are some lessons learned that we would like to share.

The **agile methodologies** are the best solution for this type of project, since we do not know what will happen with the raised challenges. They have allowed us to adapt, evolve and change the parts or technologies necessary thanks to their flexibility. Don't be afraid to make changes to proposed options and look for alternatives. It is also necessary to have the collaboration and the involvement of all the team in each iteration to advance all in the same direction.

A **multidisciplinary team** brings different points of view to solve a challenge, the contributions of both functional and technical viewpoints have contributed to improve the final solution. Also, the different points of view on the management of cultural heritage of different countries helps to enrich it.

Even if a core piece of the project are the enrichments, they have been relegated to the end of the project., **Start from the beginning**. Without a set of data to work with, they do not make any sense. The focus of the project at the beginning must be on the data.

Much of the project has consisted of research and development to find the best solution considering the large volume of data with which you work. **Learning by doing**, the best way to know if it works is to test it; and after this, test it again using different data sources.

Results and future development

After 2 years in 19 one-month iteration sprints, we have developed 7 releases and one stable minimum viable product (MPV). The developed tools allow to analyze, clean and transform data to the EDM standard. Validate and publish heterogeneous data to a normalized data lake that can be exploited as linked open data and with different data visualizations.

The Github platform is used to publish software releases under a MIT open source license, manage user requests, process contributions and also publish related documentation and a new software community based on Benevolent dictator for life model (BDFL) is waiting for users feedback.

The status of the development is to improve the data source mapping and transformation tools and focus on the enrichment module. On the other hand, more users of the platform are expected to help grow the community.

References

[1] Ariela Netiv & Walther Hasselo, ECHOES - cooperation across heritage disciplines, institutes and borders (IS&T, Washington, 2018) pg. 70-74

[2] **All the code** (version 1.4)**, specifications and documentation are available on the GitHub** page of the project: *https://github.com/CSUC/ECHOES-Tools.*

[3] Lluís M. Anglada & Sandra Reoyo & Ramon Ros & Ricard de la Vega, "Doing it together spreading ORCID among Catalan universities and researchers" (ORCID-CASRAI Joint conference, Barcelona, 2015)

[4] Victor Pasqual, The navigation system of ECHOES. Report 2018.

[5] Anisa Rula & Andrea Maurino & Carlo Batini, "Data Quality Issues in Linked Open Data". (Part of the Data-Centric Systems and Applications book series, DCSA, 2016)

[6] Europeana Data Model (EDM) *https://pro.europeana.eu/resources/standardization-tools/edm-documentation*.

[7] Duke. A tool to find duplicates. *https://github.com/larsga/Duke*

[8] Frank van Ham & Adam Perer, "Search, Show Context, expand on demand": Supporting Large Graph Exploration with Degree-of-interest. http://perer.org/papers/adamPerer-DOIGraphs-InfoVis2009.pdf

Acknowledgements

We wish to acknowledge the great work provided by Gerard Suades and David Fernandez, colleagues who have been part of the CSUC development team. It is also important to emphasize the implication and imperceptible work of Walther Hasselo, Olav Kwakman and Anna Busom, specially each month at the "cookie" meeting. Finally, we would like to thank also the help provided by the data visualization expert Víctor Pasqual, of OneTandem company.

Authors Biography

Ricard de la Vega *is the Computing and Applications Department Manager at Consorci de Serveis Universitaris de Catalunya (CSUC). He received a bachelor's degree in Software Engineering from the Polytechnic University of Catalonia (UPC), a master's degree in Computer Science from the Open University of Catalonia (UOC), a master's degree in Business Innovation and Entrepreneurship from the Pompeu Fabra University (UPF) and a postgraduate course of Big Data and Analytics from the UPC. He is interested in data related topics (big, small, open, fair, LOD, interoperability, searching, machine learning, visualization, preservation, etc).*

Natalia Torres *is the Computing and Applications Department project expert leader at Consorci de Serveis Universitaris de Catalunya (CSUC). She received a bachelor's degree in Systems Engineering from the Polytechnic University of Catalonia (UPC), a master's degree in Computer Science from the University of Catalonia (UPC). She is interested in interoperability, visualization and preservation.*

Albert Martínez *is a software engineer in the Computing and Applications Department at Consorci de Serveis Universitaris de Catalunya (CSUC). He received a bachelor's degree in Computer Science from the University of Catalonia (UAB). He is interested in big data, visualization and machine learning.*