ECHOES – cooperation across heritage disciplines, institutes and borders

Ariela Netiv & Walther Hasselo; Heritage Leiden; Leiden; The Netherlands

Abstract

ECHOES is an acronym for Empowering Communities with a Heritage Open EcoSystem.
That is exactly what the result of the project should be: an open system which can be used by (groups of) heritage organizations to connect data on very diverse types of heritage objects and information objects related to heritage.

Why ECHOES?

The first ideas for ECHOES resulted from the foundation of Heritage Leiden. Heritage Leiden is an amalgamation of a regional archive, an archaeological service, an expert centre for listed buildings in the region and a museum. The different disciplines have different collection management systems. In addition, there are several databases containing information about research and data extracted from archives and newspapers. All this information is searchable on the website via one search box, but for the results you have to go to the different databases. So, our first problem is that information on the heritage of Leiden and its surrounding area is stuck in silos of information.
Our second problem is that you have to fill a text box to find the information you might be looking for. The different databases use different terminologies and sometimes you have to be very specific in your search request. This is fine for genealogists and historians, but less practical for people who are interested in the history of the region, but do not have a very specific question. Some people just want to know if there is information on a specific location or period.
The third problem is that a lot of information on Leiden and its region is kept by the University Library of Leiden University, City Museum De Lakenhal, Science Museum Boerhaave, the National Museum of Antiquities, historical societies, and the 'dike board' Rijnland, the organization protecting us from drowning. Most of these organizations have their own searchable websites, but those sites contain a lot of information (University Library!) that is totally irrelevant for a person interested in local history.
And if one has a specific question one has to go through 7 or 8 websites to find answers.
Our fourth problem is the diversity of our user groups. They range from students doing historical research to architects and building historians looking for material concerning the restauration of a listed building, to genealogists to political scientists using our digitized newspapers. These groups have very different needs and skills.
Last, but not least, our collection management systems, our website and storage system all come from the same supplier and we want to end this vendor lock in.

Design principles

Thinking of a solution to these problems we decided that the first requirement would be that collection management systems, a middle layer and the website(s) should be separate entities that can work together but are independent of one another. We needed a scalable solution that provided three things: easy-to-use modules to convert all sorts of heritage data to one format, a middle layer that would deliver quick results to a query and a number of widgets that could be used in various websites to present the results.
The design principles we laid down were as follows:
1. User-centred design. We hired someone whose only job for the last year and a half has been to do research into current use of our websites and the problems people encounter. This is being done through interviews, mouse-stats, etc.
2. Linked open data where possible. Not all our data are suitable to be transformed into linked open data, but everything that is stored in a database is.
3. Sustainability in storage and infrastructure. For storage that means PID's, low-energy storage, safeguards against bit-rot and a host of other things. For the infrastructure that means flexible modules, so that a change in one part of the infrastructure can be effected without changing the whole.
4. We store data once, and then only if we are the source responsible for curating these data. So the database fro listed buildings should use the pictures stored in the picture database. We store the geographical coordinates for a listed building, but link those coordinates to a map for addresses that is used nationally.
5. Integration: there should be unity in access (people, places and times), data management and work processes.
6. Using the power of users. We use crowd sourcing wherever possible and the results of the work of the crowd should be easily added to our collections.
7. Intuitive interfaces. People of all sorts of skills and interests should be able to use the sites without a lot of explanation.
8. Data visualization for searches and meaningful results.

Setting up the project

After thinking out the basic project we did a market consultation to see what was available in the marketplace and to get an idea of the costs. The market consultation delivered several interesting results: the whole of what we wanted was not available, though several companies had developed components that might be useable.

Some companies had literally no idea of what we were asking and came up with solutions that were not an answer to our questions. And, last but not least, getting what we wanted was going to cost a lot of money, probably in the region of 1,000,000 to 1,500,000 €.
That meant we had to look for partners for this project. Initially we looked for partners in the Netherlands and Belgium, because they speak the same language. When it became clear that lots of institutions could contribute in kind, but that finances in the heritage sector were tight, we opted to try for European funding. That led to more partners in more countries and a bigger project due to European requirements. In the end we put in a funding request for a conglomerate of 16 businesses, universities and heritage institutes, in six countries: The Netherlands, Belgium, Sweden, Latvia, Spain/Catalonia and Croatia. When the bid for funding failed, we decided to see if we could set up a smaller version with some of the stakeholders. Tresoar, a heritage institute from Frisia in the Netherlands, the Generalitat (government) of Catalonia, the Deputácio (provincial authority) of Barcelona and the Concorsi de Serveis Universitaris de Catyalunya (CSUC), a technical university in Catalonia, joined us, Heritage Leiden, to form a new coalition. CSUC is the main developer.

A leaner, meaner ECHOES

To reach the original goals of the project the new consortium first did an analysis based on the bid for European funding. A few additional design principles were formulated:
1. Re-use of open source software that had been developed in the meantime by European projects like LoCloud.
2. Use of an open source triple store
3. Choice of a data model that could handle all types of heritage information
4. The whole of ECHOES should be open source. This a requirement for Spanish public bodies anyway.

Technical choices and specifications

CSUC performed an in depth technical analysis[1], which was the basis to start the development.

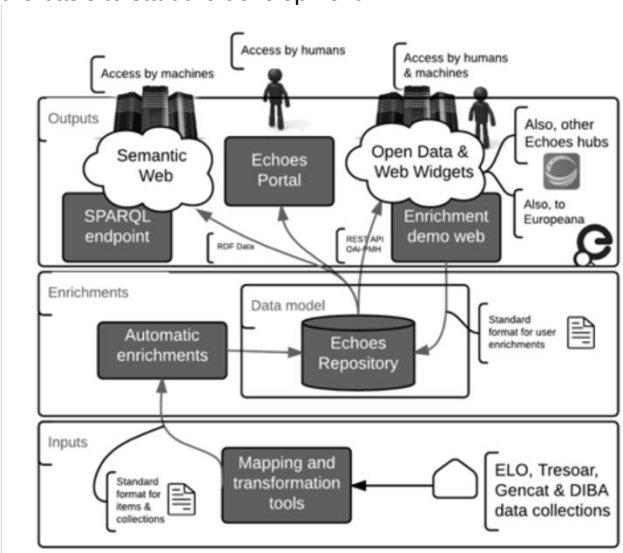

Figure 1 ECHOES architecture

In this analysis a couple of important choices were made. The most important choice was to choose EDM[2] as the data model. EDM as a data model is not only versatile enough to store the wide variety of information ECHOES has to be able to handle, it also has the advantage that information stored in an ECHOES repository can be harvested by Europeana. The following technical architecture was designed:

The development process

It was decided that this project would be done as an Agile project, working in sprints of one month. Basically the development can be split up in for parts:

Part one

This part handles the mapping to EDM and transformation of the data to Linked Data (triples). The basis of the mapping module is the core library from the LoCloud mapping tool Mint.

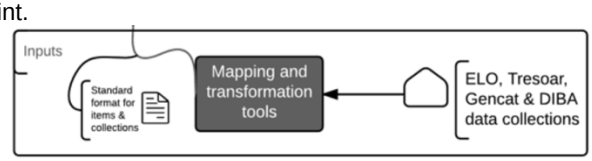

FIGURE 2 The mapping proces

Part two

Part two arranges the development of the Data Repository. Having quite some experience with Apache Fuseki (DSpace), CSUC chose this as the basis for the repository. But as soon as we started transforming and ingesting data it was clear that DSpace was not the best choice. If we wanted to use the power and possibilities of LOD we would need a triple store. We decided to install three different ones so we could do some testing on performance, but also see how they connect to websites, what other tools were available etc. The triple stores tested were: Virtuoso, Marmotta and Blazegraph. After quite extensive testing Blazegraph came out as the best choice.

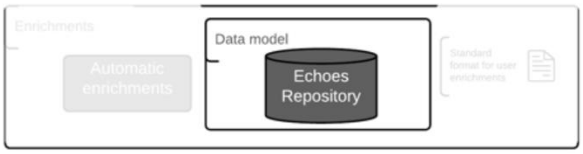

FIGURE 3 The ECHOES data repository

Part three

Part three is all about enriching the data. Enrichment will be developed as a two lane street. One lane will be about enriching data in an automated way, mapping information against sources like Geonames, DBpedia, Semium Time and ontologies like the Getty AAT.

The other lane will be manual enrichment, mainly by the end-user.
If data is shown on a website, tools will be developed to enable users to enrich the information.

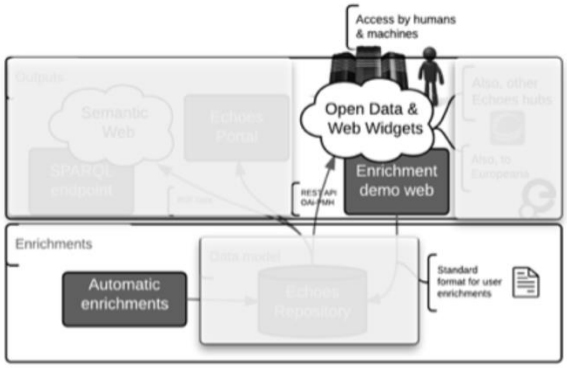

Figure 4 The enrichment process

Part four

Part for is about data retrieval and visualization.
Echoes will not only show metadata related to records, the project portal will include visualization tools that will help users understand and recognize the subject of the content shown in the item. There will be three different options to export the contents of each Echoes instance:

- Echoes Portal: A web portal with search and browse capabilities of the contents of the data repository. Additionally, other widgets and data retrieval and visualization tools can be added. This is a window for the users to interact directly with an instance of Echoes.
- Sparql endpoint: A tool for exporting the data of the project in an RDF format to the semantic web, accessible by machines.
- Open data: All the data of the repository can be exported in different kinds of formats and technologies (REST API, OAI-PMH) so that other interested parties, portals, websites, and repositories can question or harvest the data.

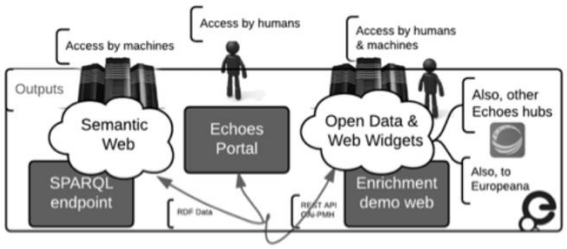

Figure 5 The data retrieval and visualisation

Lessons learned

The developers have been working on this project now for 14 months and along the road some import changes to the original plan have been made.

:

The first one was the decision to skip Apache Fuseki and use a full fledged triple store instead[3].
The second important decision was to take a step back and redesign the mapping tool. At first there was a “one size fits all” solution for the mapping. But the more varieties of data we tried to import, the more it became obvious that we actually needed something more versatile.
The developers came up with a modular approach to this part of ECHOES. A new mapping tool was developed, that consists of 5 parts:

1. Metadata schema definition
2. Morphia core
3. Parser core
4. Recollect core
5. Validation core

Figure 6 The modular approach to mapping

As simple a change as this seems, it already has proven to be a very important one, since this gives you the possibility to import almost any form of data, with only a small amount of labour needed to develop a custom mapper/validator.

Importing data from Heritage Leiden and from Tresoar has been pretty straightforward, since Dutch institutes tend to use standardized data models for various collections (Dublin core for images, A2A for family data, EAD for archival info etc.). It was when we started to ingest data from the Catalan partners that the modular approach really became vital. In the data from Generalitat and from DIBA we found different metadata on one and the same object. The challenge now was to come up with one record that would combine all the information from both institutes. The following schema shows clearly how this was done

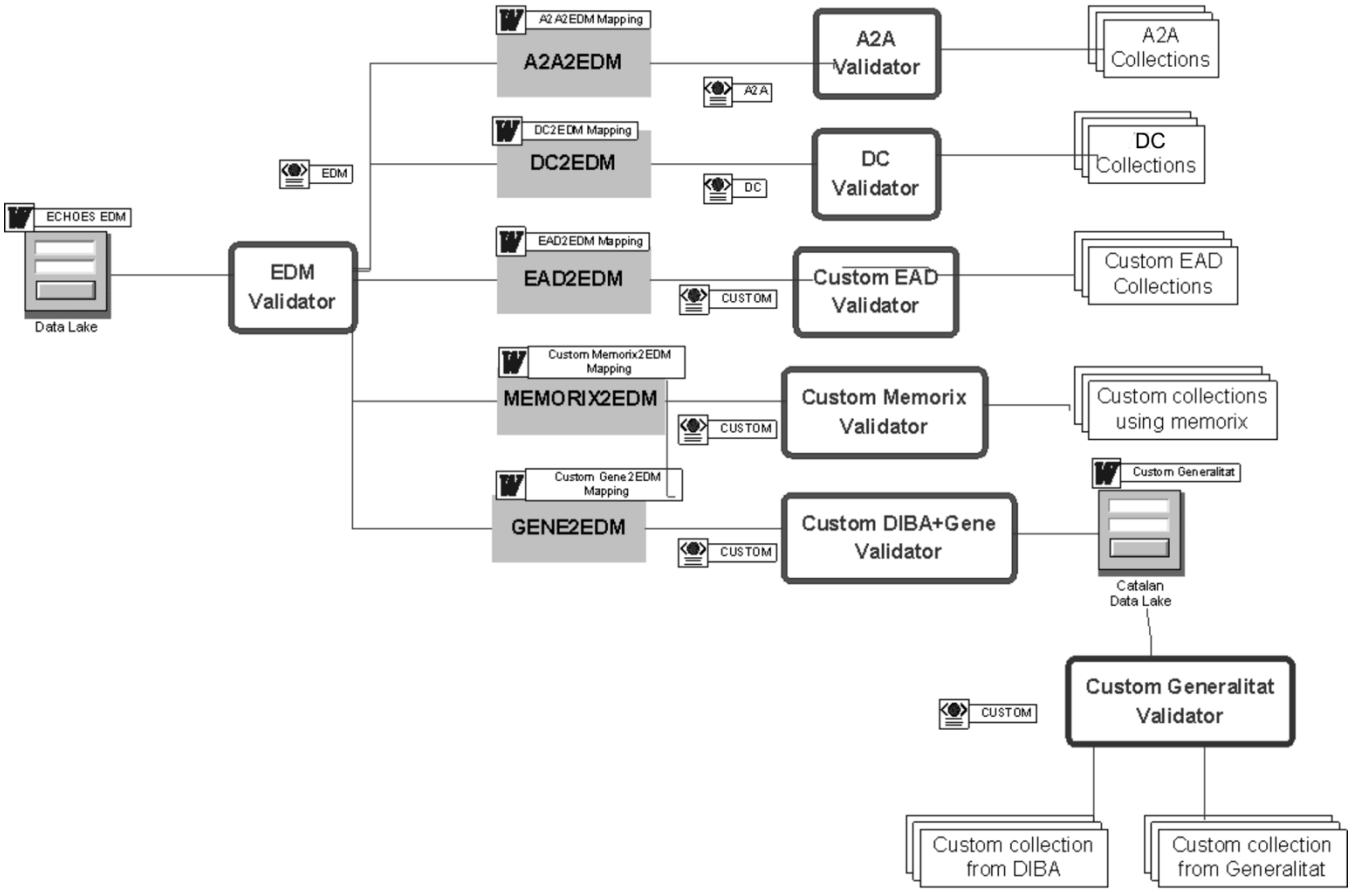

Figure 7 The workflow for mapping and transforming data sets

Using the specialised validator and mapping tools gives you the possibility to map almost everything to a single data model (EDM) and validating the data again, after mapping, with an EDM validator ensures the data is homogeneous.

ECHOES prototype in 2018

The project planning for the development runs till September 30 2018. At that time we will have achieved at least the following goals:

- The possibility to transform data from various institutes and various data models to one uniform repository where the data will be stored and accessible as linked data based on the EDM data model.
- The automatic enrichment of the data by using sources like Geonames, DBpedia, AAT etc.
- The possibility to access and query the data through a text search box on a simple website.
- The possibility to access and query the data through the use of a map as a search interface
- The visualization of the linked data of at least one type of information

Further development of ECHOES

When this project ends on September 30 2018, not all we would like to do will have been done, so on October 1st 2018, the next ECHOES project will start. In this project the main

focus will be on unleashing the power of Linked Data and data visualization. Using the automatic enrichment will also mean opening up new ways of questioning at the data. Since it is not common yet to offer cultural heritage information as linked data, there is also a huge task to design the new ways of accessing the data in as user friendly a way as possible. Companies developing back office systems for cultural heritage institutes have shown interest in the project as well, so maybe part of the further development will also include cooperation in that area.

What is the difference between ECHOES and Europeana, DPLA etc.?

Wonderful projects like Europeana et. al. are vary valuable for bringing cultural heritage to a large audience. For smaller institutes or local societies it is, however, almost impossible to add their information to Europeana.

ECHOES takes a more local approach, making it easy to add data and publish this on a website, bringing together information based on either a region or theme. At the same time it makes it easy to add this information to a platform like Europeana.
For people interested in local history the use of ECHOES means they do not have to plough through the enormous amount of data available in Europeana. And for a local institute or historical society it means their data can be added easily to the European treasure trove that is Europeana.

Use cases for ECHOES

From the beginning we had several use cases in mind: Firstly, the access to all historical collections in Leiden in a Shared Heritage Leiden Portal. Through ECHOES, data from the University Library of Leiden University, City Museum De Lakenhal, Science Museum Boerhaave, the National Museum of Antiquities, historical societies, and the 'dike board' Rijnland can be searched and shown in one place. This means that somebody doing research on Roman Leiden will find books from the University Library, archaeological dig reports, Roman pottery, books and photo's from Heritage Leiden, swords and masks from the National Museum of Antiquities and pottery and dig reports from the Archaeological Society in one place.
A selection of these data can be used by Leiden Marketing to illustrate the Limes tourist route along the border of the Roman empire, which runs through Leiden.
This combination of local collections in one place, without having to modify the Collection Management Systems, in a low cost way through a local installation of ECHOES, is possible for locations like Friesland, an area of the Netherlands where they speak Friesian, or the mountain villages in the Pyrenees in Catalonia. As long as managers of local collections have a systematic way of describing their assets, even if it is just in an Excel sheet, they can contribute to a local platform. Because it is very simple to convert their data, even amateur collectors can join in the fun.
The local data, because it is presented as Linked Open Data, will be enriched as never before, giving context to local history. And the aficionado of local history will be able to search by marking a location on the map, instead of having to think about search terms.
The visualization tools will show connections that haven't been noticed before, e.g. the visualization of the number of agricultural workers in the databases of Heritage Leiden gives a good idea of the spread of the bulb cultivation business in the area.
ECHOES can also be used to bring together very specific information on a theme. In 2020 Mayflower400, the sailing of the Pilgrims to America, will be commemorated. Information on the Pilgrims is kept in archives, libraries and museums in England, the Netherlands and the United States. They came from England, fled to the Netherlands in 1608 and lived in Leiden from 1609. From 1620 onwards a number of the refugees left for America, accompanied by adventurers from England.
Anyone who wants to see their whole story needs ECHOES to bring it together.

References

[1] A copy of this report can be found on the ECHOES GitHub space: https://github.com/CSUC/ECHOES-Tools/tree/master/background-information

[2] EDM:Europeana Data Model

[3] Test reports on the performance of the 3 different triple stores are also available in the background information section of the ECHOES GitHub space as mentioned under reference [1]

Biography

Ariela Netiv received her MA in mediaeval history (1988) from Leiden University. She trained as an archivist at the National Archive School (1990). She joined the city archive of Leiden, becoming its director in 1999 and developing it towards a regional heritage centre, Heritage Leiden. She has served on several national committees, e.g. the Arts Council and the Prins Bernhard Culture Fund. Her focus is on finding new ways to bring heritage to the public.

Walther Hasselo is an Oracle DBA and certified Prince2 project manager. He is co-author of the guidelines for Open Data policies in Cultural Heritage He has given numerous presentations on new ways of publishing cultural heritage information at varies international conferences. He works at Heritage Leiden as project manager e-Heritage, in which capacity he is also project manager of the ECHOES project.