Archives Portal Europe: a new frontier in digital space

Go Sugimoto
Netherlands

1. Lift-off

There has been a big trend within the cultural heritage sector over the last decade to develop web portals at every level, ranging from a local museum community to an international library consortium. Especially in Europe, this phenomenon is distinct as the needs of the portals are identified due to the barriers which complex layers of institutions and administrations tend to build across the continent. So, what is the story so far for archives in the 21st century of the digital universe?

The European national archives have looked at ways to develop an archives portal since the 90's. The eContentplus programme of the European Commission (EC) gave them the opportunity to realise this, and as a result of this instrument of the information society and media policies that the EC promoted under the name of i2010 and the Digital Agenda, the APEnet project[1] was born. It is a project to create a common access point to European archival descriptions and digital collections. It also delivers data to Europeana[2], the EC flagship project to build a cross-domain cultural heritage portal. As such, the APEnet project played an extremely important role, acting as a data aggregator of the archival domain in Europe.

After years of dedication and effort, the APEnet project celebrated the public debut of Archives Portal Europe (http://archivesportaleurope.net) (Figure 1) in January 2011 as a pilot installation to showcase the potential of descriptions and digital collections from 17 European national

Figure 1
Archives
Portal Europe

archives. It contained over 8 million descriptive units and gave access to more than 750,000 digital archival objects, which was noticed and welcomed by the European archives community. Since then the website has continued to add content and implemented new functionalities. In March 2012, the APEx project[3] was launched, replacing the APEnet, with 28 national archival institutions plus ICARUS[4] on board. Subsequently we are now looking at truly staggering numbers: the total content of 23 million descriptive units and 40 million digital objects in them. The success of the portal has a long story behind the scenes.

This article addresses a part of that story, describing in each chapter one of the challenges we have confronted and how we have dealt with it. It also describes the value and the scope of the project, particularly focusing on the technical aspects of building a web portal through which users can explore the richness of archival material across Europe.

2. Data engineering - standardisation of standard(s)

Building the Archives Portal Europe means collecting data about archival materials from our content providers to make them searchable. These data are called holdings guides at a general level and finding aids at a more precise level. It is important to note that, in the case of digitised archives, the APEx project only acquires the descriptions and not the digital objects (e.g. digital files of scanned documents) themselves. Thus, the portal points to the physical locations of archival materials and serves hyperlinks to the original websites of the content providers. In this way the system is de-centralised and Internet traffic is diverted to the content providers.

A survey was conducted to compare the different kinds of finding aids existing in European national archives. The outcome led to the conclusion that EAD[5] is widely in use as a standard, however, it is adopted in many different ways. Consequently further standardisation was needed to create a common profile for the realisation of a fully interoperable web service.

After intensive discussions about the "standardisation of the standard", the project defined the so-called apeEAD schema. In order to use the aggregated data in a machine-processable way and to display it in a user-friendly manner, this schema is a vital ingredient. It is the basis for the portal service, and the Data Preparation Tool (DPT) which the APEx project offers for the content providers is engineered to transform all local EAD data into apeEAD data.

In addition to the finding aids which deliver information on the archival materials, an archival landscape is created, offering a general overview on the archival institutions participating in the project, first ordered by country, then following each country's specific administrative or geographical organisation. It is linked to the EAG which includes information about institutions such as visiting address and opening hours, and on services and facilities offered onsite. The finding aids, holdings guides, and archival landscape are the three layers of the entire data structure which form union finding aids, the core of the Archives Portal Europe framework.

Towards the end of the project, some other standards such as EAC-CPF[7] and METS[8] will be included in the portal. EAC-CPF is known as an authority file of agents including persons,

families, and corporate bodies, which deliver more contextual information for the archival research. It will bring added value to the Portal, demonstrating connections from archival materials to people, and from people to other people, for example. METS will make it easy for archivists to manage a large volume of digitised objects in finding aids with the possibility of maintaining proper rights information. The APEx project is committed to be an active player in the standardisation in the archives domain. The details of the use of those standards can be found at our project website[9].

3. Not rocket science - search engine and data presentation

The quality of search engine and display is one of the most significant outcomes of the project. It is about how end users interact with the portal and look for information. However, implementing a good search engine for a vast amount of diverse data is not an easy task. It is not just a matter of indexing provided data. It is necessary to consider various requirements and options for the configurations of indexing, harmonisation, standardisation, refining, and reordering of the search results. There is a lot of invisible work behind the portal. Our developers have specified an optimal mechanism in order to guarantee both high-speed access to huge hierarchical data and to improve data visibility on small computer screens.

The APEx project members believe that the portal demonstrates the proof of our concept for archival data navigation, retrieval, and presentation. We have created a two-display system which facilitates easy and quick access to the essential information. The users first see a results page and from there a second display shows each detailed results (Figure 2). In addition, two tab options in the first display enable the users to switch between two views of the results: a simple list view (Figure 3) and a context view (Figure 4). The former is a quick and familiar way to display the results all at once with the possibility of refining and reordering, whereas the latter shows the same results grouped following the archival landscape and presented within their archival context. Advanced search allows the users to search and explore the content from different angles. They can select the collections or holdings to be searched. They can also navigate the content within the whole data structure, divided by countries first, secondly by content providers, and finally by their holdings. This navigation method provides an excellent opportunity to explore and browse the richly structured content and easily assimilate, without even noticing it, the basis of archival knowledge.

The flexibility of the functionalities and interface will stimulate the use of the portal by a wide range of users from archivists and experienced researchers to school students and teachers. The current interface is a hint of potential and the work is still in progress. It will be improved toward future releases in such a way that the portal meets the needs of the end users as much as possible. In fact, at the time of transition from APEnet to APEx, it was recognised that end-user evaluation is a high priority task. Evaluation activities are on-going and planned over the space of three years. They will be able to identify the user profiles and their needs; thus the portal will be developed in close communication with the users.
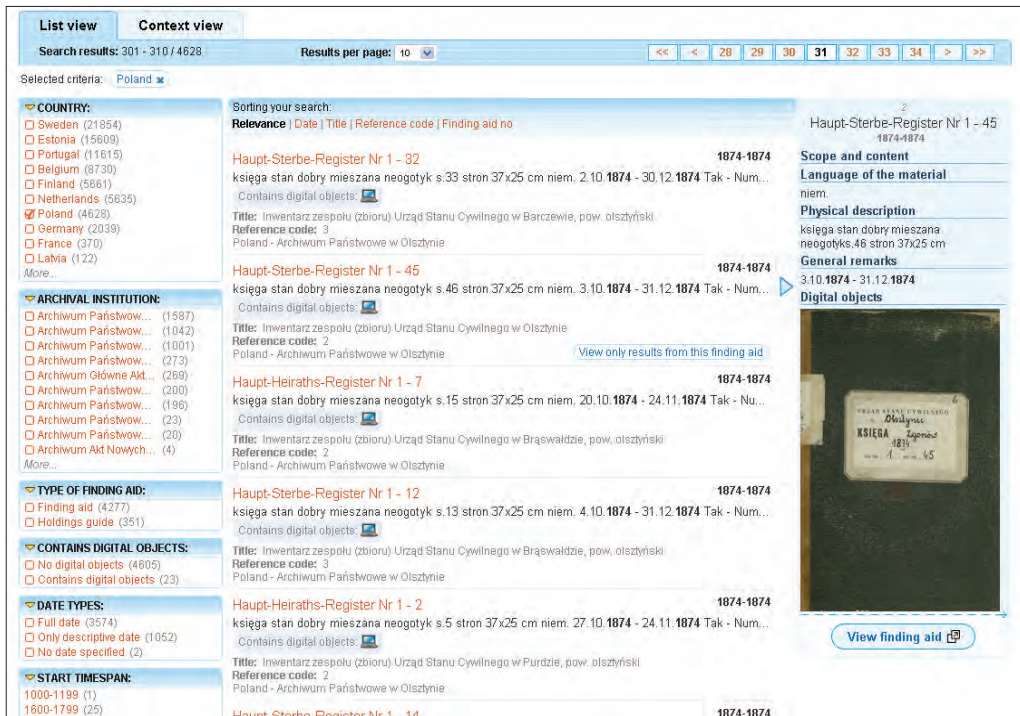
Figure 2
The second display
of search results

Figure 3
List view
of the search
results
(first display)

Figure 4
Context view of the
search results
(first display)

4. Docking – reinforcing Europeana

There is no doubt that synergy with Europeana is another priority of our project and therefore we have a close liaison with them on a political as well as a technical level. Data contribution from APEx enforces the raison d'etre of Europeana since archives are – together with museums, libraries and audio-visual collections – the most important pillars of Europeana aggregation.

This is why we have developed the Europeana conversion functionalities within the standalone DPT and our centralised Dashboard web service (see next chapter) to convert apeEAD into Europeana Semantic Elements (ESE), the metadata format which Europeana requires for the data submission. This data mapping and conversion involved a challenge, which originated from the different approaches to cultural heritage objects and metadata. While Europeana focuses on object-based metadata about digital objects only, Archives Portal Europe provides access to the descriptions of objects that are not digitised yet and at the same time uses a top-down overview of data – via the finding aids – to deliver the context information of the archival objects.

Our mapping specifications struck a balance between the two approaches by supplying information on digital objects only, but with as much context information from the higher levels of the hierarchically-structured finding aids as much as possible. Consequently, Europeana data conversion was implemented in the DPT and Dashboard, supporting the best possible mapping between two completely different data schemas.

This situation has changed recently in the course of technical progress. During the last few years there has been an enormous collaborative effort by both projects, realising the establishment of Europeana's new schema, called Europeana Data Model (EDM). To supplement the simple ESE schema, the EDM is designed to handle EAD's complex hierarchical data model and demonstrates the possibility of capturing the essence of archival descriptions. Although there are still some remaining concerns about fine-tuning of the mapping between apeEAD and EDM as well as a solution to the tricky display issues of archival data in Europeana, we are confident that the EDM overcomes the problems of ESE and gives the users a richer experience to explore semantic data networks. This constructive teamwork with Europeana will continue, but as a first milestone, APEx project is now on the verge of finalising the EDM conversion in order to maximise the semantic interoperability with Europeana. A flexible mapping implementation with a user-friendly dialogue interface ensures the most interoperability with the least effort. The function also fits perfectly in the chain of data workflow as implemented in the Archives Portal Europe back-end. In addition, an OAI-PMH[10] repository is deployed so that Europeana can regularly harvest our EDM records. As a result, millions of records will be potentially ready to be delivered to Europeana.

## 5. Running low in orbit - sustainable autonomous infrastructure

Cultural heritage organisations have always struggled to find a sustainable business model for their web services. Everybody involved asks the same question over and over again: how can we build a sustainable portal without spending a lot of money to maintain its quality? It is hard to find an answer, but one solution is to make operating costs as low as possible.

As we see, the central aggregation approach is popular and widespread. Europeana and The European Library[11] are good examples of this concept. They request cultural heritage institutions to submit their metadata in order to make the aggregated data searchable in their portals. However their teams need to coordinate the data delivery and to process the data centrally. They inevitably increase the running costs as the number of the providers and the data sets grow. Even using common standards and automated workflow, there is an enormous effort to aggregate, manage, and maintain data at a large scale. Not that this mechanism costs technically, but it is probably even more costly for the administrations to organise the content provider network in Europe. It is said that over 2000 institutions are involved in Europeana aggregation.

The APEx project has a different approach. We offer our content providers full control over the management of their data, thus minimising the operating cost of the central administration. For this purpose we developed a centralised web data management system in the back-end of the Archives Portal Europe - the Dashboard (Figure 5). It serves as an all-in-one application which has all necessary functions for data management: uploading, validating, converting, normalising and indexing/publishing data, and delivering them to Europeana. The content providers are given access to this secured system where they can decide themselves what they would like to do with their data. On the other hand APEx technical team still centrally monitors and manages the Dashboard and portal as a whole. Consequently the implementation of the Dashboard not only

Figure 5
Autonomous
backend application
– The Dashboard
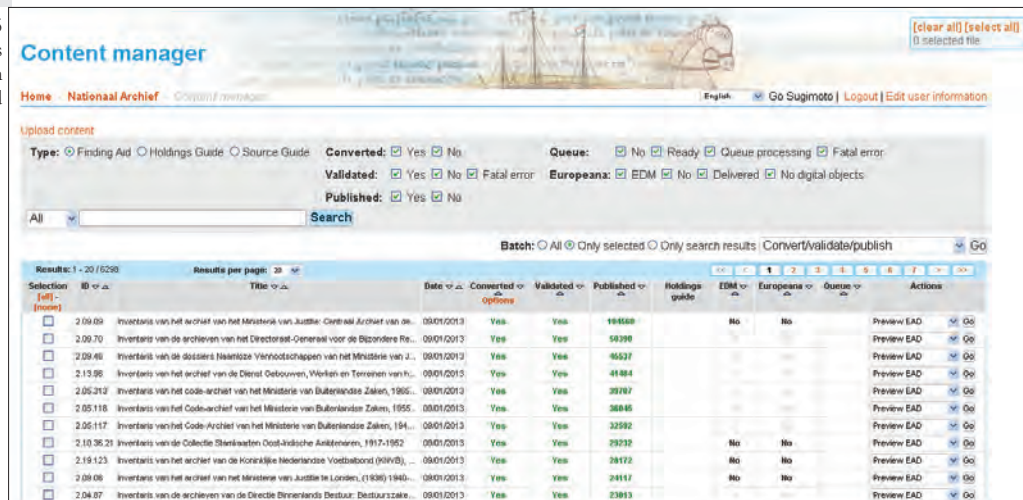
helps to maintain the portal cost-effectively, but also accommodates the autonomy of the content providers.

In addition, the above-mentioned DPT is available as an Open Source software, equipped with extended functionalities for local use. It allows the content providers to prepare data locally and to have even more control over the data they provide to the Archives Portal Europe. Needless to say, this also helps in reducing the performance cost of the central server. The whole data management workflow is optimised as much as possible to minimise a seemingly big effort of the content providers. As a result, the combination of the local tool and the centralised management system enables us to evenly distribute the workload and responsibility and simultaneously to keep control of the entire data management. Although this strategy creates the goodwill incentive of knowledge transfer and distribution of responsibility, the inevitable use of the sophisticated tools could be a hurdle for participation in APEx and technical expertise largely varies in each content provider. It is, therefore, our mission to deliver good documentation and training. Workshops are also held in order for the users to have hands-on opportunities to work with their data and for us to communicate with them for feedback.

6. Landing - Digital Renaissance and Archives Portal Europe

The APEnet and APEx are the first serious attempts of the European archives community to work together to build a large-scale Internet gateway (Figure 6), and should be able to provide many opportunities. Archivists and researchers can discover new insights into our history, while the citizens may enjoy seeking information about their ancestors and teachers will have an abundance of material for school education at their disposal. Up to now the APEx project has been mostly operated by national archives because of their expertise; however, it is planned to expand its network activities more to regional and local archives in the coming years. As a matter

Figure 6
The APEx project
website

of fact, national-level coordination has started to flourish as we organise workshops and conferences in different regions of Europe. A large amount of content is expected to be delivered to the portal to the extent that truly pan-European research can be carried out. This joint-effort of European archives will continue and the Archives Portal Europe aims to become the place-to-be for access to European archival materials.

It is also envisaged that the content of our portal will be ubiquitously (re-) distributed in the future through different channels such as OAI-PMH and Linked Open Data, and will be accessible via various emerging devices including smartphones and tablet computers. We are looking into the potential of Web2.0 to implement personal research spaces, feedback tools, and crowd-sourcing. Synchronising the open data movement happening in the cultural heritage

sphere, Archives Portal Europe is also eager to develop a system platform for the content providers to easily publish and share their data with a wider audience. When the end-users get more engaged with the open data and enrich it, public knowledge will increase. Although some data issues such as copyright, security, integrity, and authenticity may arise and need to be resolved, new formation of knowledge and social life will be waiting for us when the portal becomes a first-class resource that captures the diversity and unity of the European culture.

In 2011 the European Commission published a report: The New Renaissance – Report of the 'Comité des Sages'[12] describing the importance of access and preservation of digital cultural heritage and the recommendations for the coming years. The scope of the APEx project corresponds to the intention of the report; therefore, it will benefit from the current strand of cultural heritage digitisation to a great extent. The Renaissance revolutionalised human life in many ways. The boom of a cultural phenomenon symbolised by masterpieces of great artists elevated the social, economic, and political structure of Europe to an entirely new level. Similarly, our website has the potential to bring about a digital revolution for the European culture and contribute to a significant impact on our ever-changing modern society.

In this digital space-age we are now landing on a new planet. A new frontier and horizon are in front of us. We have to build a shelter to store data about our past and present, to cultivate the land, and harvest new data. We not only preserve them, but also prepare them for consumption and distribution. Data can be consumed in ways we have never known before: they can be re-used and enriched; they can re-create art and culture, thereby boosting the economy. Undoubtedly we will also face many more challenges when we proceed, but we are happy to be a pioneer in taking them onboard. We are proud of our achievements so far and will continue working for the exciting time to come.

1. APEnet project website (http://www.apenet.eu)
2. http://www.europeana.eu
3. APEx project website (http://www.apex-project.eu)
4. http://icar-us.eu
5. EAD (Encoded Archival Description)
6. EAD (Encoded Archival Guide)
7. EAC-CPT (Encoded Archival Context - Corporate Bodies, Persons, Families)
8. METS (Metadata Encoding & Transmission Standards)
9. http://www.apex-project.eu/index.php/outcomes/standards
10. OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting)
11. http://www.theeuropeanlibrary.org
12. http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf