Conference “Imatge I Recerca”
Girona, Spain
November 16-19 2004

Towards a sustainable infrastructure for access to distributed digital historical photographic collections. Lessons learned from the EVAMP project[1].

Authors:
René van Horik (rene.van.horik@niwil.knaw.nl)
and
Rutger Kramer (rutger.kramer@niwi.knaw.nl)

NIWI-KNAW
Netherlands Institute for Scientific Information Services
P.O. Box 95110
1090 HC Amsterdam
The Netherlands

Abstract

Increasingly historical photographic collections are converted into digital formats and made accessible on Internet. This paper reports on the activities carried out within the framework of the EVAMP project to realise a sustainable infrastructure for access to distributed historical photographic collections. Background information on the EVAMP project is given.
The main part of this paper describes the most important components of an information system, the refined EVA system, which main function it is to provide “Open Access” to distributed digital historical photographic collections. The realisation of the components is based on a number of assumptions that are elaborated on in this paper.
The four main components of the EVA system are: (1) support for a standardised communication protocol created by the Open Archives Initiative, (2) metadata structured according to the standardised Dublin Core metadata element set, (3) an Open Source application development platform, and (4) a software module to enable multi-lingual searching.
This paper contains essential technical background information on the components of the refined EVA information system.

[1] The authors would like to thank Douwe Zeldenrust for remarks and suggestions on the text of this paper.

Introduction

Increasingly historical photographic collections are converted into digital formats. On Internet an enumerable amount of digital image collections can be found. Several guidelines and standards are available to guide the conversion process of analogue original photographs into digital surrogates[2]. The best way to digitise a photographic collection is determined by the characteristics of the collection (format, number of items, copyright issues) and available resources to digitise photographic items (budget, skills, equipment). The online access facilities provided by Internet applications are the most important catalyst behind numerous initiatives to digitise primary visual source material.

This paper reports on the realisation of an information system, accessible on the Internet, to get access to distributed digital photographic collections by using open standards, open source software and a dedicated service provider that provides multilingual search and retrieval facilities. This information system has the name “EVA System” and is based on a prototype system with the same name. A main part of the activities to realise the refined EVA system are carried out within the framework of the international EVAMP project (EVAMP: “European Visual Archives Market Validation Project”).

This paper consists of four parts. First, background information is given on the EVAMP project. One of the work packages of the EVAMP project provided resources to enhance the existing proto-type information system giving access to digital historical photograph collections. The features of this existing prototype information system are discussed resulting in a motivation for the refinements that are considered as necessary in order to realise a more sustainable information system. Common access, or interoperable access, to digital historical photograph collections via a central portal is the most essential characteristic of the EVA information system. The second part of this paper discusses modes of interoperability. Based on the “harvesting” interoperability model the prototype EVA system is refined. In part three of this paper the main components of this refined EVA system are presented and discussed. The last part of this paper consists of a conclusion.

This paper has a rather technical approach towards the realisation of an infrastructure for transparent access to distributed digital historical photograph collections. Of course also economical, social and organisation factors will determine whether a suggested approach will be accepted and applied by an extensive international user community. These important issues are not covered in the paper. The fact that a formal organisational body representing and executing “EVA related” activities does not exist yet is an important reason that the growth and extension of the services was minimal. Again, the ambition of the EVAMP project is to solve this by taking the initiative to found an organisation that supports and maintains a sustainable service concerning the access to digital historical photographic collections. The intention of this paper is to inform keepers of historical photographic collections on the advantages of using the technical approach presented in this paper, based on standardised protocols and services.

1. The “European visual archives market validation project” (EVAMP)

By definition a project is delimited in time. Once the project goals are achieved a project is finished. The goal of the EVA project (“European Visual Archives”) carried out under the umbrella of the EU Info2000 program in the period from December 1998 until February 2001, was to inform keepers of historical photograph collections on issues related to the digitisation and dissemination of digital images based on historical photographs. Within the framework of the project a number of research reports were published and a prototype information system was developed that provides access to two digital historical photograph collections. The name of this information system is “EVA system”[3]. The EVA system showed the potential of web-based information systems providing access to digitised historical photographic collections. The results of the temporarily EVA project did not lead into a lasting service containing more

[2] See for instance “Guides to quality in visual resource imaging” (Research Libraries Group, July 2000) <http://www.rlg.org/visguides>

[3] Both the reports and the prototype information system created within the framework of the EVA project are accessible via <http://www.eva-eu.org>. This website and system is not maintained since 2002. Probably the enhanced EVA system developed within the framework of the EVAMP project will be accessible via this address.

collections and extended services. The implementation of this requires additional resources. The EVAMP project (acronym for "EVA Market Validation Project") provided the means to assist in the transformation of the provisional results of the EVA project into sustainable services related to the digitisation and access of historical photographic collections of public funded organisations. EVAMP, partly funded under the EU eTEN program[4], started in February 2004 with an intended term of 18 months.
The EVAMP project consists of a number of work packages. One work package of the project consists of a market research in order to map the interest of organisations holding historical photographic collections into services and products related to the digitisation and dissemination of photographic items. Another important work package of the project concerns the development of a business plan upon which sustainable products and services can be based on a cost-effective basis. The EVAMP project also contains a work package directed towards the refinement of the existing prototype information system as it was developed by the EVA project. This work package has the title "refinement of the prototype" and consumes about 20% of the total resources available within the EVAMP project. This refinement is based on experiences gained with the prototype EVA system, the outcomes of the user survey[5] and the initial discussions on the EVAMP business plan that is currently under development. This implies that some assumptions can be made concerning the requirements of a refined EVA system that not necessarily will be confirmed by the outcomes of the business plan.

2. Modes of Interoperability

Basically the EVA system, both the current prototype and the refined system developed in the EVAMP project, are rather simple search and retrieval applications as they can be found on Internet in abundance. The purpose of the EVA system is to provide access to digital images of historical photographs that belong to the holdings of a number of individual collections. The multi-lingual search facility is considered as a very important advanced feature of the EVA system and will be discussed further on in this paper. It is assumed that the way the collections of individual organisations are made available in a common interface very much influences the willingness and feasibility of organisations to participate in an initiative to give access to its digital image collection. The challenge is to convince organisations of the quality of the system resulting in the allocation of resources to join the initiative. It is further assumed that the less an organisation has to do to connect its already existing digital metadata and images to a joint initiative, the lower the threshold is to participate. This "re-cycle" principle for digital content is an important aspect of the proposed approach discussed further on in this paper.
A big obstacle for transparent access to a number of digital archives is the fact that many organisations use different proprietary technologies that do not allow for interoperability. This turned out to be a complicating factor during the development of the EVA prototype. Three approaches towards interoperability can be distinguished:

- Federating. In this model a group of organisations agrees that their services will be built according to certain specifications. These organisations together form a federation and have to communicate intensively in order to comply with the common standards. Depending on the complexity of the required common guidelines, this can be very costly. Federations have small but dedicated memberships. Some projects apply the federating model of interoperability to achieve transparent access via Internet to a number of collections that agree on common standards. In the Netherlands the project "Beeldbank Noord-Holland", see <http://www.beeldbank-nh.nl> provides access to about 50 collections, most of them kept by individual organisations. Currently the system contains about 100.000 items. Another example of a federating system is "Bildarchiv Austria", see: <http://www.bildarchiv.at/> currently bringing together four collections close related to the National Library of Austria. A federating system giving access to image collections across institutes

[4] For more information on the EU eTEN program, see: <http://europa.eu.int/information_society/programmes/eten/index_en.htm>
[5] The report on the EVAMP user survey in PDF format can be found at: <http://www.evamp.org/documents/EVAMP_31103_WP2_D2-2_SurveyReport_v1-2(final).pdf>

based in a number of countries could not be found by the authors. It is true that in essence the prototype EVA system applies the federating approach towards interoperability, but the experiences in the EVA project made clear that the maintenance of the data and the costly intensive consultation are the main obstacles to make this model the basis for a successful sustainable cost-effective service.

- Harvesting. Participants make some small efforts to enable some basic shared services, without specifying a complete set of agreements. The participants maintain full control over the digital sources. This approach towards interoperability is gaining importance in the digital library community. The Open Archives Initiative (OAI)[6] is providing the protocol that can be used to implement the harvesting approach towards interoperability. This model of interoperability is used as a basis for the refinement of the EVA system. Further on in this paper more arguments are given in favour of this approach.
- Gathering. If their is no cooperation among organisations at all a base level of interoperability is achieved by gathering openly accessible information, e.g. by web search engines. The quality of these services is rather poor and information hidden in databases is not reached. The "show image" option available in the Google web search application shows the poor quality of the gathering interoperability approach for digital images.

Four years ago the prototype EVA system probably was one of the first online information systems that gave access to digital historical photograph collections kept by public funded organisations. Today an uncountable number of applications can be found ranging from basic electronic exhibitions to sophisticated information systems. The EVAMP project has the ambition to provide products and services covering the complete conversion and access chain and this will be worked out in the business plan of the project. The business plan of the EVAMP project will contain federation services by offering guidelines, products and consultancy. But organisations interested in providing common access via Internet to its digital historical photograph collection, should be free to use any solution and still be able to join the EVA system without much effort. The harvesting mode of interoperability based on the open access paradigm is the best way to achieve this. The remaining part of this paper elaborates on the components that are part of the refined EVA system that enables interoperability between distributed collections of digital historical photographs.

3. Components of the refined EVA system

In this section of the paper the most important components of the refined EVA system are discussed. The refined EVA system is based on the existing prototype EVA system[7]. The function of the refined system will be merely the same as the prototype system - namely a search and retrieval application with some multi-lingual search facilities, as well as the possibility to order reproductions - but the architecture of the refined system will fundamentally be changed.
Before the most important components of the refined EVA system are discussed in detail, first attention is paid to the data formats that are supported by the EVA system. Two types of data formats are used by the EVA system: a data format for the digital images and a data format for the documentation or metadata.

1. Data format for the digital surrogates of the historical photographs. The digital images should be available in a data format that can be rendered by an Internet browser, such as Netscape Navigator or Internet Explorer. The de-facto standard at the moment is the JPEG image format[8]. In principle the specifications of the JPEG image, regarding dynamic range and number of pixels is up to keeper of the collection. The digital image must be identified by a unique internet-address in the

[6] The Open Archives Initiative, see: <http://www.openarchives.org>

[7] A description of the prototype EVA system can be found in: R. van Horik, "Archives and Photographs: the 'European Visual Archive' Project (EVA)", Cultivate Interactive, issue 3, 29 January 2001. <http://www.cultivate-int.org/issue3/eva/>

[8] More information on the JPEG image file format can be found at: <http://www.jpeg.org>

form of a so-called Uniform Resource Locator (URL). A URL is a compact string representation of the location for a resource that is available via the Internet[9].

2. Data format for the documentation or metadata of the objects in the EVA system. The metadata must be formatted in the XML format[10] (XML: eXtensible Markup Language). XML is an information interchange format, developed as a standard by the World Wide Web consortium. The format is application independent, both human and machine-readable. The standard is extensible because it does not have any pre-defined mark-up tags. Within the EVA system the XML files do have to meet some requirements, concerning the structure of the XML file and the tags in the XML file. This issue is explained further on in this paper.

In the next sections of this paper four important components of the refined EVA system are described. These four distinguished features of the refined EVA system are the main building blocks of a sustainable infrastructure for access to distributed digital historical photographic collections. First, the implementation of "Open Access" is described that facilitates the efficient dissemination of content based on the harvesting mode of interoperability. Next, the role and relevance of a standardised metadata element set is described. The third building block of the refined EVA system consists of an Open Source development platform and the last building block concerns a third party software component that makes it possible that the documentation of the digital images can be stated and accessed in six European languages.

3.1 Open Access by means of OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting)[11]

The goal of the "Open Archives Initiative – Protocol for Metadata Harvesting" (OAI-PMH) is to supply and promote an application-independent interoperability framework that can be used by a variety of communities that are engaged in publishing content on the Web. The OAI-PMH protocol permits metadata harvesting. Increasingly organizations are developing dissemination systems based on this protocol, as they are aspiring to unlock access to their collections in a simple and low-maintenance manner, and to make interchange of metadata between multiple heterogeneous archives possible.

The OAI-PMH protocol effectively removes the dependencies on system architecture and metadata compatibility. Archives opening up their collections using OAI-PMH merely have to install a Data Provider module acting as a uniform access point to the metadata. A Data Provider maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata. The Data Provider is a passive software component connecting the metadata repository to the Internet by the use of the XML data format and the Hypertext Transmission Protocol (HTTP)[12] protocol, and can be implemented to offer Open Access using multiple metadata formats. Data Providers are systems that support the OAI-PMH as a means of exposing metadata.

As soon as the Data Provider is realised, it can be accessed and harvested by Service Providers. A Service Provider issues OAI-PMH requests to Data Providers and uses the metadata as a basis for building value-added services. This means that another system can connect to the Data Provider and download the metadata it needs in one of the available metadata formats. The downloaded metadata is stored and indexed by a Service Provider that will offer the actual end-user access to the metadata. One Service Provider can harvest as many collections as it sees and offer access to all of these repositories at once, making it seem the end-user is accessing one large collection. The refined EVA system can be considered as a Service Provider where as the organisations that make their digital images available to the EVA system can be considered as Data Providers.

The OAI-PMH is a communication protocol or language with only six permitted verbs that are allowed to be used by the speakers of the language. The verbs are given in table 1, together with a short description of the function of the verb.

[9] More information on addressing in identifying objects on Internet can be found at: <http://www.w3.org/Addressing/>
[10] For more information on XML, see: <http://www.w3c.org/xml>
[11] For more information on OAI-PMH, see: <http://www.openarchives.org>
[12] For more information on HTTP, see: <http://www.w3c.org/Protocols/>

Verb	Function
Identify	Give a adscription of repository
ListMetadataFormats	Give a list of metadata formats that are supported by repository
ListSets	Give a list of sets that are defined by repository
ListIdentifiers	Give a list of unique OAI identifiers contained in repository
ListRecords	Listing of N records
GetRecord	Listing of a single record

Table 1: The 6 OIA-PMH verbs

The verbs “Identify” and “ListMetaDataFormats” and “ListSets” are related to repositories that are part of a Data Provider. The other three verbs are the actual harvesting verbs: they make it possible to locate an individual resource within a repository.

When the Service Provider contacts one of the Data Providers for the first time, it will use the “Identify” verb. The Data Provider will respond with a data block containing relevant information about the organization hosting the Data Provider and Data Provider itself.

After the Service Provider has received and interpreted the identification information it will ask the Data Provider which metadata formats it supports by issuing a “ListMetadataFormats” verb. This yields a list of metadata formats, their validation schema locations and a metadata prefix. This is a unique code that identifies the metadata format in the repository; i.e. whenever a Service Provider wants information in a specific format, it should specify this format using the appropriate metadata prefix. In case the metadata is structured according to the Dublin Core metadata element set (DCMES) the metadata prefix is “OAI_DC”. DCMES is discussed further on in this paper.

When the Service Provider has made its choice from the available metadata formats, it can start harvesting the records. Two verbs are available to perform this task. The “ListRecords” verb will yield all the records in the repository and the “ListIdentifiers” verb will yield all the OAI-Identifiers available in the repository. An OAI-Identifier identifies a unique record in a repository and can be used to retrieve a single specific record using the “GetRecord” verb.

The Service Provider will eventually have retrieved all of the records from the repository that it can index and offer access to for end-users. To ensure that the harvested collections are up to date, it should check with the Data Provider regularly for changes by issuing either a “ListRecords” or a “ListIdentifiers” verb with the last harvesting date as a parameter. The Data Provider will then return all of the records or identifiers that have been changed or deleted since the last visit. The scenario using the “ListRecords” verb is shown in the figure 1.

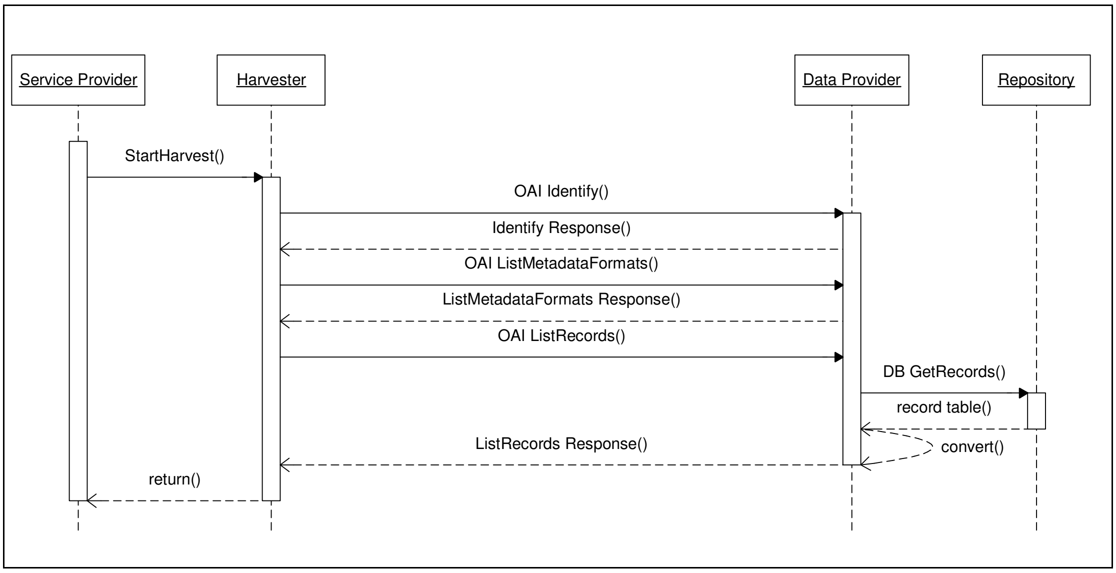


Figure 1: Scenario for a simple harvesting session using the “LISTRecords” verb.

The establishment of a Data Provider that can be harvested by the refined EVA system is done in the following way. The digital images that represent the historical photographs are placed on a Web server. A URL uniquely identifies each individual image. The metadata or documentation concerning the digital image is stored in XML format according to the syntax that is explained in the next section. A simple process that makes it possible that a Service Provider can access the metadata is installed on the server that hosts the metadata. This is the Data Provider. A number of public domain Data Providers are available[13].

It is the intention of the business plan created by the EVAMP project to define services, such as consultancy, training and manuals to assist in the installation of a Data Provider. It is foreseen that in the near future increasingly information system will support OAI-PMH as a standard feature. In the digital library community "Open Access" is already more or less a standard function.

The Data Provider can be seen as an interface between the Internet and the repository. It effectively is, or runs as parts of a web server, and handles requests that come in from a harvester. It always has two connection points, as shown in Figure 2

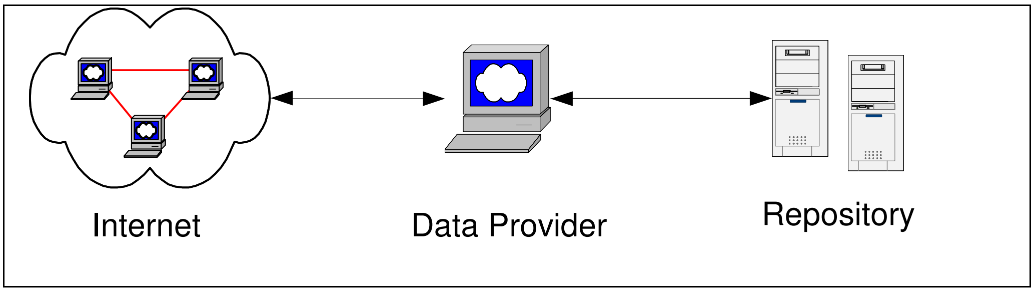


Figure 2: Data Provider architecture

The connection on the left ties the data provider to the Internet. This is usually implemented by using third party web server software, such as ASP and TOMCAT. Data Providers are available on a wide variety of web server architectures. There are also Data Providers available than have an embedded web server system and as such offer a turnkey solution.

Creating the connection of the Data Provider with the repository is more intricate and depends on the repository's architecture, implementation and data format. The simplest implementation is based on the metadata records being stored as separate XML files on the file system. In this case, whenever the Data Provider needs to return metadata records it only has to iterate over the XML files, parse them into the OAI result and return it to the harvester. A more common repository implementation is based on a database management system, such as Microsoft Access or Oracle. Data Providers can access these databases by creating a connection with the help of the ODBC or JDBC standard.

Other configurations are also possible, for instance database systems that are not ODBC or JDBC compatible. In these cases, Data Providers will have to be custom-made or extensively altered in order to work with the repository system. If this is not possible for some reason, one can always resort to the first approach: metadata in XML format stored on the file system, by exporting the records from the database and reformatting them to XML. Almost all database management systems have an XML data export function.

The Data Provider is responsible for converting the metadata into different structures or formats. This is usually done by defining crosswalks to the alternative formats, and performing these crosswalks at the time of dissemination. A crosswalk is a mapping between metadata schemas. In a crosswalk it is decided e.g. that the data element "Title" of metadata schema X can be mapped with the data element "Caption" of metadata schema Z.

Defining a crosswalk and implementing a Data Provider that performs the mapping requires a good understanding of the semantics of the data elements that are part of a metadata element set. Crosswalks can be defined using XSLT style sheets[14] and standard Open Source software is available to perform the crosswalk conversion fully automated.

[13] An overview of registered Data Providers are available at: <http://www.openarchives.org/Register/BrowseSites.pl>

[14] For more information on the transformation and presentation of XML formatted data with the help of Style sheets, see: <http://www.w3c.org/Style/XSL/>

3.2 Metadata according to the Dublin Core data element set (DCMES)

The second important component of the refined EVA system is a set of standardised data elements, known under the name Dublin Core metadata element set (DCMES)[15]. A data element is a unit of data for which the definition, identification, and permissible values are specified by means of a set of attributes. Organisations engaged in the digitisation of historical photographs use a wide range of sets of data elements. These data elements often describe a number of resources such as the analogue original photograph, the digital surrogate as well as the scene visible on the image. The EVA project learned that only after intensive consultation and communication it is possible to reach agreement on a set of data elements for a common access system to digital historical photograph collections. It is foreseen that this can dim the readiness of organisations to join such an initiative.

This is the main reason that it was decided to use the DCMES as the main metadata standard in the refined EVA system. DCMES has the status of an official ISO Standard as ISO15836:2003. DCMES consists of 15 data elements that are stated in table 2. OAI-PMH supports any set of data elements, but DCMES is defined as the preferred metadata format. The main rationale behind the development of DCMES is to serve as a set of data elements to be used for resource discovery on the Web. In principle DCMES is applicable for the resource discovery on Internet for any object.

Data element	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content of the resource.
Subject	A topic of the content of the resource.
Description	An account of the content of the resource.
Publisher	An entity responsible for making the resource available
Contributor	An entity responsible for making contributions to the content of the resource.
Date	A date of an event in the lifecycle of the resource.
Type	The nature or genre of the content of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A Reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content of the resource.
Rights	Information about rights held in and over the resource.

Table 2: The 15 data elements of DCMES

In order to lower the threshold as much as possible for organisations to provide the refined EVA system with content it is up to the organisation that holds the photograph collection to decide which DCMES data elements are used and how to interpret the meaning of the data elements. It is not difficult to imagine the wide range of alternatives to interpret and apply the data elements. The DCMES data element “Date” e.g. can be interpreted by individual organisations as a global time period by mentioning the begin year and end year of this period or as an indication of a specific day. Common agreement on the exact definition of the data elements would of course result in more homogenous metadata resulting in an information system of higher quality, but for the refined EVA system DCMES is used in an unqualified way.

For two data elements of DCMES within the framework of the refined EVA system the usage is obligatory. The data elements “Description” and “Identifier” must be populated with data for the refined EVA system. The data element “Description” should contain as much as possible data on the content of the image, because the Service Provider indexes the data in this data element and uses this data element for the full text retrieval. The data element “Identifier” must contain a link to the digital image as it can be accessed with the help of an Internet browser. Figure 3 contains an example of a record that is formatted in the XML data format according to the DCMES standard.

[15] More information on DCMES can be found at: <http://www.dublincore.org>

```
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/">
        <dc:title> lang="en"
            Open Spaces Committee on a visit
        </dc:title>
        <dc:description>
            Metropolitan Board of Works: Parks, Commons and Open Spaces Committee on a visit, 1884
        </dc:description>
        <dc:date>
            1884
        </dc:date>
        <dc:creator>
            Greater London Council
        </dc:creator>
        <dc:identifier>
            http://www.imageserver.org/image/lma/l1307AR.jpg
        </dc:identifier>
        <dc:publisher>
            London Metropolitan Archives
        </dc:publisher>
        <dc:keywords>
            Parks
        </dc:keywords>
</metadata>
```

Figure 3: An example of a record in XML format using DCMES data elements

Both the EVA project and the EVAMP project made clear that the compilation of homogeneous metadata to be used in a common access information system is not as easy as it seems in first instance, even if the rather general DCMES set of elements is used as a common format. In most situations a conversion between a local used set of data elements into a DCMES compliant structure has to be carried out. This process can be compared with the crosswalk activity described in section 3.1 of this paper.
Table 3 and table 4 illustrate the heterogeneous situation concerning the data elements used for the description of digital historical photographic collections. The data elements used by a number of arbitrarily selected online accessible information systems that provide access to digital historical photographs are compared with each other. Table 3 gives information on the organisations of which the online access systems are selected.

Country	Name of system	URL of web interface	Type of Institute
New Zealand	TimeFrames, National library of New Zealand	http://timeframes.natlib.govt.nz	Library
Austria	Bildarchiv, National Library of Austria	http://www.bildarchiv.at/	Library
United Kingdom	Photograph search, Imperial war museum	http://www.iwmcollections.org.uk/qryPhotoImg.asp	Museum
France	Musee de la photographie (departement de l'essonne)	http://www.photographie.essonne.fr	Museum
Netherlands	Beeldbank Nationaal Archief	http://beeldbank.nationaalarchief.nl	Archive
Spain	Fototeca Sevilla	http://www.fototeca.us.es	Archive

Table 3: Some organisations providing online access to digital historical photographs[16]

Table 4 gives an overview of the search field labels that can be found in the online access systems of the institutes stated in table 3. It can be assumed that the data elements used in the web interface give an indication of the way digital data objects are enriched by data elements and to what extend these data elements can be mapped to the DCMES standard. Intentionally, the labels of the data elements are not translated or explained in any detail to make clear that only an insider who is aware of the backgrounds of the individual collections is able to determine how the local used data elements should be mapped to the DCMES data elements.

[16] The Websites were consulted and active in August 2004.

All information systems have both a full text search facility, often indicated as “simple search”, to access the collection of digital images, and an advanced search facility. The advanced search facility contains a number of search input fields that can be used to create a detailed search command. These search fields are related to the data elements that are used to document the items available in the system. The search fields of the web-based systems are stated in table 4. The labels in table 4 are listed in the same order as visible in the web interface. Most web interfaces contain background information on the meaning of the search fields. But none of the systems mentions a standard set of data elements as a reference for the compilation of the search fields.

TimeFrames, National library of New Zealand	**Bildarchiv, National Library of Austria**	**Photograph search, Imperial war museum**	**Musee de la photographie**	**Beeldbank Nationaal Archief**	**Fototeca Sevilla**
Title	Suchbegriff(e)	Subject	Catégorie	Beschrijving	Nombre
Year	Person	Photographer	Sout catégorie	Dag/Maand/Jaar	Autor
Name	Schlagword	Colour/B&W	Procéde	Periode vanaf/tot en met	Localization
Subjects	Bildnummer	Period	Auteur	Trefwoord	Año entre ...y ...
Iwi/Hapu	Klassifizie-rung	Photograph Number	Auteur secondaire	Collectie	Tipo
Place	Medientyp	Collection Number	Titre	Fotograaf	Formato
Descriptive notes	Technik		Legende	Fotonummer	Serie
Image type	Institution		Sujet	Commentaar	Fotógrafo
Reference number	Jahr		Lieu de prise de vue		Fecha entre ... y ...
			Négatif á l’origine		Soporte
			Support		Sección
			Theme iconographique		
			Sous theme iconographique		
			Epoque		
			Date		
			Usage de destination		
			Usage connu		
			Recherche		

Table 4: Search fields in web based access systems to digital photograph collections

Table 4 makes clear that a wide range of data elements is used for the description of digital historical photographs. Based on the name of the data elements in several cases a relation between the local used data elements and the data elements of DCMES is evident. But several other data elements can only be connected to DCMES data elements by an insider that knows about the backgrounds of the local used set of data elements. It is also possible that the local used data elements cannot be linked to any data element of DCMES. The translation of local used data elements to DCMES data elements is a matter of semantic interpretation of the data elements and this cannot be done automatically, but requires the input of a person that is familiar with the meaning of the local used data elements. The importance and relevance of DCMES is that it is a broad, but generally accepted standard making it worthwhile to carry out the crosswalk in order to improve the interoperability between distributed digital historical photographic collections. The translation of local used data elements to data elements of the DCMES standard is an important basis of the refined EVA system.

3.3 Open Source development platform

A third important component of the refined EVA system concerns an Open Source development platform. Open Source means that the computer source code of the Service Provider as well as the Data Provider is accessible for everybody. The source code can be re-used provided that other developers gain credit for their work done. The relevance of an Open Source development platform for the refined EVA system is based on the following three assumptions:

1. Usage of Open Source software will improve the trust and durability of the information system. A proprietary system would require specific measurements in order to guarantee that in the future the system will not become obsolete or surprise the users with new conditions and requirements.
2. Open Source software used for the refined EVA system will potentially attract a wider user community, because users of a diversity of specific software products only for a small number of its functions – namely basic online resource discovery - use an Open Source solution.
3. Lastly, the use of Open Source software development ensures that the users of the EVA system will not be dependant on one or more specific software manufacturers for maintenance or possible changes to the system, which will enhance the flexibility of the system.

The Service Provider of the refined EVA system is developed with the help of the i-Tor platform[17]. i-Tor is an Open Source technology that enables you to create websites. They may be straightforward web pages, or information extracted from a database or from a Data Provider. i-Tor can also be used to make modifications: the creator of a web page can manage it directly on the site, either alone or in collaboration with others. Users can search all of the information that is linked to i-Tor. Support for the OAI-PMH has been added to the I-Tor platform and consists of both Data Providing services as well as Service Providing services. A number of high quality commercial software products are available that can be used to realise online access to digital historical photograph collections. These digital asset management systems contain much more functions than the current refined EVA system. Examples are functions to create extensive tailor-made documentation, including thesauri and specific data entry quality control mechanisms. Next to that advanced image processing functions might be available in a number of products, e.g. the possibility to zoom in and facilities to protect the digital image. A couple of years ago only a few commercial digital asset management systems contained an interface that was accessible with an Internet browser. Today this is more or less a standard feature. Indicators in de Digital Library community learn that in the near future also the support of OAI-PMH will be a standard feature. One of the partners of the EVAMP project is a company specialised in products and services in the field of digitisation and dissemination of pictorial collections of museums, libraries and archives. Their digital asset management product supports the OAI-PMH already making it very easy to set up a Data Provider[18].

3.4 Multilingual search facility

A fourth important building block of the refined EVA system is an improved multilingual search facility. Acting on an international scale it is obvious that metadata of digital images of historical photographs is available in a number of languages and that users of the information system originate from several countries speaking a wide range of languages. Thus, a facility to search in a multilingual document collection, a process usually referred to as Cross-Language Information Retrieval (CLIR) is of great relevance for the EVA system. The most difficult challenge of CLIR is considered to be sense disambiguation as it determines the most

[17] For more information on i-Tor, see <http://www.i-tor.org/en/>. The source code of i-Tor can be found at: <http://sourceforge.net/projects/i-tor>. A review of a number of Open Source repository systems, including I-tor can be found at: <http://www.soros.org/openaccess/software/>.

[18] This company is Pictura Imaginis and the digital asset management system they develop is called "Memorix". See: <http://www.pictura-dp.nl/index.php?lang=eng>

suitable translation term to be used in a particular context. The English term “bank” e.g. can both refer to a financial institution and to a waterfront.

The prototype EVA system contained a function to search in the collection in Dutch, German and English[19]. This multilingual search function was based on a proprietary algorithm that required man-made wordlists or lexicons as input parameters for the translation algorithm. The lexicons contained all relevant terms extracted from the descriptions of the images. Moreover, the lexicons must be available in all the languages supported by the information system. Also a so-called “expansion list” was created manually containing specific terms and related narrower terms. In the expansion list e.g. the term “church” is labelled as a narrow term of the term “building”.

Despite the fact that the query translation module of the prototype EVA system turned out to perform satisfactory, it was decided that the refined EVA system required a replacement of the existing multi-lingual search function. The main reason for this was that the creation of the lexicons is very labour-intensive and thus too expensive to be used in the refined EVA system. The Dutch, German and English lexicon and expansion list that is used in the CLIR of the prototype EVA system contain about 6.000 terms. It took the full time attention of an editor for about 10 months to create the three lexicons and expansion list.

Within the EVAMP project research was carried out to find a CLIR system that requires a minimum of manual interaction and that supports more languages than the three supported languages of the prototype EVA system[20]. In line with the other modules of the refined EVA system a preference existed for the development of an Open Source solution. For the development of an Open Source CLIR system dictionaries are required that contain terms in the languages supported by the information system. It turned out that only for a limited number of languages public available dictionaries are available: English, French, German and Dutch. Public funded research projects have created open sources for multi-lingual purposes, but they are only free for academic research. For other usage large sums of money is required. A proof of concept made clear that a linguistically inclined programmer is able to build a solid modular Open Source CLIR, but the initial costs to build the CLIR as well as the already mentioned lack of dictionaries caused a shift of attention toward commercial available CLIR services.

A number of CLIR services and products were investigated, compared and analysed. It was concluded that the Dutch Company Irion[21] was able to provide the most optimal solution to realise a CLIR as part of the refined EVA system. Irion exploits the linguistic heritage of the EU-funded project “TwentyOne”[22]. The services and products of Irion are based on a semantic network build up from a number of lexical resources. Languages supported are Dutch, English, Spanish, French, German and Italian. This is only a fraction of the number of languages spoken in Europe, but services and products suitable for the EVA-system that support more languages could not be found. Next to that the CLIR framework build by Irion is able to support more languages in the future. Among the functions of the semantic network of Irion are the recognition of synonyms, the handling of compound words and phrases.

The metadata harvested by the EVA Service Provider is indexed by the CLIR system of Irion and processed by their semantic network. A search string of a user of the refined EVA system in one of the six supported languages is mapped with the multilingual index and in order of relevance descriptions of images are presented that meet the requirements of the user.

The main benefit of using the third party solution for multi-lingual searching is that the process is fully automatic.

The licensing policy of Irion is very flexible and modular resulting in a situation that the costs for a CLIR serviced by Irion are much lower than the development of an Open Source solution would have been. The licensing structure of Irion can easily be adapted to future situations taking the volume of metadata to be indexed into account as well as the number of queries, the refreshing ratio of the indexes and the fine-tuning of the system for a specific language. Hopefully the number of languages supported by the system will improve in the future increasing the number of potential collections to join the EVA initiative.

[19] A detailed description of the multilingual search facility of the prototype EVA system can be found in the publication mentioned in note 6.

[20] Stefan Rijnhart, ‘Finding a replacement for EVA’s Query Translation Engine’. Internal EVAMP report (February 2004).

[21] For more information on Irion, see: <http://www.irion.nl>

[22] Goal of the project TwentyOne was “development of a multimedia and multilingual information transaction and dissemination tool”. The project ran from 1996 until 1999. For more information on the project, see: <http://dis.tpd.tno.nl/twentyone/>.

4. Conclusion

A survey carried out within the framework of the EVAMP project made clear that institutes that hold historical photograph collections find digital conversion and Internet access to photograph collections important[23]. A business model aiming at the realisation of sustainable services for digitising historical photographs and providing access to digital images is currently under development. This paper presents a number of assumptions that are based on experiences and knowledge gained in both the EVA and EVAMP projects. Based on these assumptions an information system is realised that can be used to set up a common access system to distributed digital historical photograph collections. Table 5 contains the four most important assumptions used for the development of the refined EVA system, as well as the components of the system that are based on these assumptions. The four assumptions of table 5 are covered in section 3 of this paper.

Assumption	**Design principle of refined EVA system**	**Component of refined EVA system**
"Harvesting" is at the moment the best mode to achieve interoperability between distributed digital historical photograph collections.	Open Access.	Service Provider using OAI-PMH. (Distributed collections should act as Data Providers).
A simple, standardised metadata format is the only solution for the opening up of distributed collections that use a diversity of metadata formats.	Standard Metadata format.	Support for Dublin Core metadata element set (DCMES)
Using Open Source software for the development of the refined EVA system, will improve the trust participants have in the system as well as the durability of the system.	Open Source development platform.	Development based on the i-Tor open source technology.
A good multi-lingual search facility is important "unique selling point" for the refined EVA system. Development and maintenance of dedicated wordlists and expansion lists is too labour intensive, so application of dedicated third party solution is preferred.	Multi-lingual search facility provided by specialised service provider.	"TwentyOne" toolkit for natural language processing in a number of European languages (English, French, German, Italian, Spanish and Dutch)

Table 5: Four important components of refined EVA system

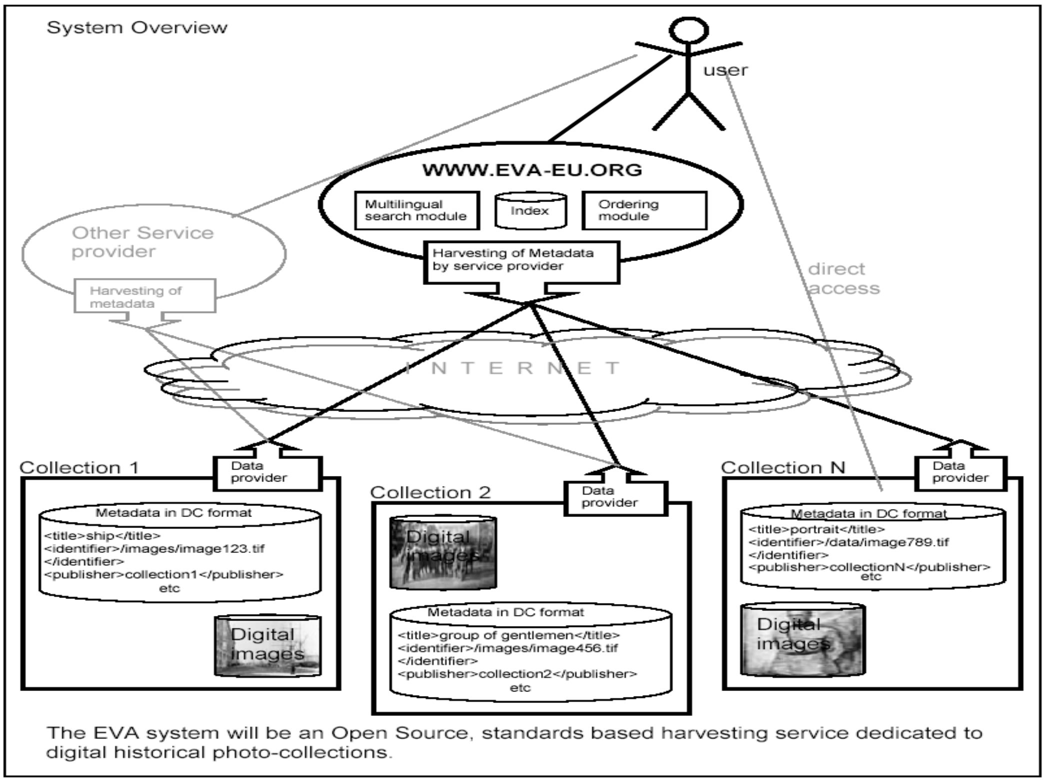
It can be stated that the technical issues to realise a portal system that gives access to distributed digital collections are relative easy to settle. The technical issues are subordinate to the quality of the content available in the system. The technical threshold for institutes to join the initiative in principle is rather low. Actually convincing organisations to join is more complicated.

Looking to the digital library community learns that after a while Open Access is supported on a wide scale, also by funding bodies. It is time now that also the archival community sees that Open Access has benefits to open up its holdings in a common interoperable interface.

It is foreseen that next to the refined EVA system a number of other Service Providers will come into existence, maybe acting as competitors to each other. Just as a book or an electronic article can be made accessible in a number of online catalogues also digital images of historical photographs can be "re-cycled" in a number of ways. It is the ambition of the refined EVA system to become one of the obvious portals to go to for common access to distributed digital historical photograph collections. The future will learn whether this ambition is realistic or not and whether the assumptions stated in this article are valid. Both the EVA and EVAMP project made clear that there is a lot of interest both with content providers and

[23] For details on the EVAMP survey, see note 4.

content users to set up a common service. This paper made clear that standards, protocols and tools are available to realise this situation. Figure 4 contains a model of the refined EVA system. The autonomous distributed collections expose metadata in XML format in the DCMES format to a Service Provider. The Service Provider (refined EVA System) harvests and indexes the data and gives access to the data via Internet.
Two specific functions of the refined EVA system are increasing the rationale for Data Providers to provide the EVA system with content. The first one is the multilingual search facility that has the benefit that the metadata can be accessed in six different languages and thus extending its user community enormously. The second advantage of the EVA system above other Data Providers is the availability of the ordering module. This module, just as the ordering module in the current prototype information system enables users to sent a request to Data Provider to produce a reproduction of an image. Organisations are able to gain some income from the system.



The EVA system will be an Open Source, standards based harvesting service dedicated to digital historical photo-collections.

Figure 4: System overview of refined EVA system