D 6.1 and D 6.2 Functional and technical design

Project Number :	**PUB 1128 EVA 25001**
Project Title :	**EVA: European Visual Archive**
Deliverable Type :	**(PUB)**
Deliverable Number :	**D.6.1 and D.6.2**
Contractual Date of Delivery :	**March 2000**
Actual Date of Delivery :	**June 2000**
Title of Report :	**Functional and technical design**
Work Package contributing to the Deliverable :	**WP6**
Nature of the Deliverable :	**report**
Authors	**EVA CONSORTIUM**

System design “EVA-System”
Workpackage 6 EVA-project ("Working Model")

Project code: PUB 1128 EVA 25001

Contents

1. Scope of the project

This project will result in a working system that disseminates information on historical photographs from various public archives. The system is called the *EVA-system*. Its main purpose is to allow end users to *discover photographic resources*.

The development of the EVA-system is part of Workpackage 6 of the *EVA (European Visual Archive*) Project. EVA is a project for the Info2000 initiative launched by DG XIII of the European Commission, responsible for telecommunications and the information market. Info2000 projects are multi-national, public-private sector partnerships that exploit public sector information. The official site of the EVA-project can be found at: `http://www.eva-eu.org`. The project code of the EVA-project is PUB 1128 EVA 25001.

As a starting point the content of the EVA-system will be a collection of 20.000 photographs of the city archives of Antwerp (SAA) and London (LMA). Both SAA and LMA are considered as representative for most European main city archives. Local traditions and perspectives are respected by trying to establish an approach that will fit in both infrastructures.

The overall project management of the development of the EVA-system is the responsibility of Telepolis, Antwerp.

2. Purpose of this document

The purpose of this document is to describe the functional and architectural design of the EVA-system. In order to achieve a design a problem analysis has been performed. The document contains the result of the internal discussions about the functionality needed in the EVA-system. NIWI acts as the editor of this document.

The information in this document is updated regularly. A "document history" statement at the top of the document provides information on the most recent adjustments. Once all project members agree on the result a fixed document will be created that will act as the execution plan for the development of the EVA system. This document will be posted on the closed listserver available via email address `EVA@nic.surfnet.nl`. The document archive of this listserver can be reached at url `http://listserv.surfnet.nl/archives/eva.html`. Subscribers have to apply for a password in order to get access to the web version of the listserver.

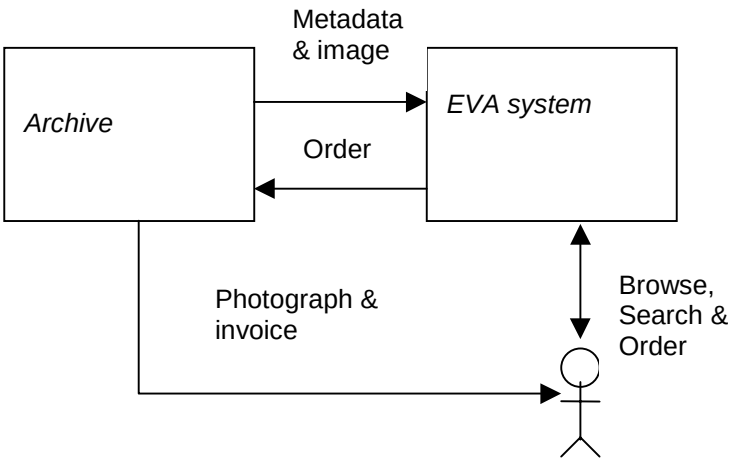
3. Basic Principles of the EVA system

The EVA-system is an information system designed mainly to provide access to individual photographs that are part of historical photographic collections. These photographic collections belong to the holdings of public archives. Via the Internet a user can get access to a catalogue of historical photographs and search and browse through the description fields. The user can view digital images of historical photographs and the user can order prints and digital images. The user can be any individual person or organisation, e.g. multimedia industry and publishers. To facilitate access to as much users as possible some multilingual functions will be part of the system.

The central information unit of the EVA-system is a description of a scene visible on a photo graph (the content) and the digital image that represents the original photo. This unit has a unique identity code in the system. The code is the key to exchange data between the central EVA-system and local information systems used by the archival organisations.

The EVA-system makes it possible for any type of user to order specific items, but the actual transaction between the user and the owner of the item (the archive) is done directly between the user and the archive. The EVA-system establishes the contact between the user and the archive. So the system communicates the wish of the user to the archive. Next to that the system will inform the user on prices, formats and shipping procedures of the item the user is interested in. In case a user is interested in the acquisition of an image (photographic print, digital print, digital image) data of the users as well as the image and output format is sent to the archive that owns the photograph. The system will facilitate the entry of personal data of persons/institutes that would like to buy an image.

Figure 1: the basic principle of the EVA system



The EVA-system does not facilitate maintenance on the metadata and images directly. The system facilitates the periodical import of descriptions and images. Mutations and additions should be done in the local “back office” system and an export/dump of descriptions and images will be imported in the EVA-system. The result of this import of “scene descriptions” and “reference images” can be that existing units are mutated or that new units are added.

For a participating archive the EVA-system can be considered as a “front-end” to its photographic collection. This front-end contains “resource discovery” descriptions of individual photographs and relatively low quality images. The EVA-system will provide information on how to create full descriptions and digital master image files. This information is the output of other work-packages of the EVA project. The EVA-system will be an important medium for the dissemination of the knowledge gathered in the EVA project.

Some parts of the EVA-system will be available in several languages. This means that labels and explanations of the system will be available in several languages. The main language, however will be English. A specific section of the search-facility of the EVA-system contains a multilingual search option. A user can enter a query in natural language and the system tries to find relevant images in the collection.

An important function of the EVA-system is to facilitate "resource discovery". This means that the descriptions of the images and photographs in the EVA-system are targeted at non-specialist users accessing the collection via the Web. Thus, the descriptions are not as detailed and extended as can be expected in an archive management system. The EVA-system does not contain images of which the copyright is not cleared. The description elements are taken from the Dublin Core (DC) standard (For an overview of all 15 Dublin Core elements see: `http://www.purl.org/dc` and EVA workpackage 3.2).

3.1 Information in the EVA-system

The information in the EVA-system consists of two items:
Digital images
Information about these images (“meta-data”)

Ad 1 Image specifications
The EVA-system will contain digital images of the original photographs of which textual descriptions are accessible. The purpose of these images is to give the user a fair impression of the original. The specifications of these “reference” images will be established in consultation with the archives. The EVA-system will contain basic discovery information on the original photograph and the digital versions of the original photograph. The system will also contain information on the specifications and prices of prints that a user can order. The more pixels and grey levels in a digital file, the higher the quality of the print.

Ad 2 Information about the images
It must be realised that the end user of the system is mainly interested in the type of information we call "resource discovery". He won't be interested in all the technical aspects of imaging. So it won't be necessary to store information on things such as tonal range.

3.2 Meta data elements

The information objects in the EVA-system have the following attributes:

ID: A code uniquely identifying the scene in the EVA-system.

Title: Brief description of the scene visible on the image / photograph.

Description: Free text description of the scene.

Photographer: Creator of the original photograph.

Date: Date. This is the date the photograph was taken (it should also be possible to supply a time period instead of a date).

Relation: Inventory number of the original photograph. A relation code with the EUAN system can be given in this attribute. The local archive has to decide on this.

Geography: Geographical keyword. This can be anything from a street name to the name of a river.

Language: The language the metadata is in.

Thumbnail: Image file name of the thumbnail image.

Ref. image: Image file name of the reference image.

Subject: A keyword term. Participating archives will supply this attribute according to their own scheme: there will be no standard thesaurus - at least initially.

Archive: Name of archive that owns the original photograph.

These attributes can be mapped to Dublin Core Elements in the following manner:

EVA-elements	**Dublin Core elements**
ID	---
Title	Title
Description	Description
Photographer	Creator
Date	Date
Relation	Relation
Geography	Coverage
Language	Language
Subject	Subject
Thumbnail	Identifier
Image	Identifier
Archive	Publisher

4. Functional requirements of the EVA system

In order to distinguish the various functions the system should be able to perform, a use case analysis has been performed. Here the *actors* (persons using the system) are related to the actions they perform described in a use case.

4.1 Actors and their use cases

Two types of actors have been distinguished, namely end users and archive employees. Both actors perform various use cases. Requirements of how the system should work are gathered. These so-called functional requirements are organised per use case.

4.1.1 General requirements

Some requirements apply to the whole system. These are presented here.

The system will provide clear and sufficient information to help the end user in operating the system.
The system will provide static pages in the following languages: Dutch, English, French, German, Italian and Spanish. The default language will be English.
General help information will be provided on the introduction screen of the EVA system.

4.1.2 The end user

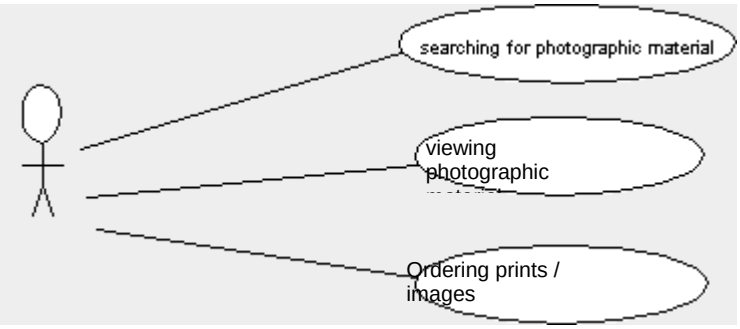


Figure 2: the use case diagram of the end user

The end user wants to search for photographic material and view the results found. These are really two different things: searching and viewing. And maybe the end users also wants to order a reproduction of a retrieved image.

Searching for photographic material

As to searching the following requirements have been formulated:
An end user can select in which collection(s) (from which archive(s)) he wants to have access to.
The system should facilitate insertion of search criteria for the following fields separately:

- keyword/subject
- photographers name
- time (period)
- geographical location

The system should facilitate full text search (on the title and the description of the image).
The system should provide the possibility to search the content using Boolean functions.
It should be possible to enter a query using terms in Dutch, German and English and yet retrieve records that contain translations of those terms.

Viewing photographic material

Another set of requirements applies to viewing.
The end user will have the option to browse through the collection(s).
Initially the records resulting from a query should be displayed as thumbnail images.
The thumbnails should function as a hyperlink to the corresponding reference image.
The number of thumbnails to be downloaded in one chunk should be limited. If a result set is large (number of retrieved records bigger than 10) it should be downloaded in chunks of 10.
The end user should have the option to choose whether he wants to retrieve the thumbnail, the meta-data or both.
The attribute Title will be displayed with each thumbnail. The reference image is accompanied by all meta data. The meta data will be in the original language.
Together with the presented visual material an intellectual property statement should be displayed.

Ordering photographic material

An end user might be interested in buying a photographic reproduction of an image found while querying the system.
The system should provide the possibility to order a retrieved photograph.
The system should contain information about:
- The fact that the system only takes the order and redirects it to the archive possessing the photograph.
- The general conditions surrounding the transaction (this information will be different for every archive): prices and details of the order procedure.

A disclaimer about the intellectual property provided by every archive separately.
Per photograph the end user must insert the following (again this is archive specific):
- purpose of usage (e.g. research, education, publication. Note the EVA-system will not cover the ordering of images used in commercial high volume productions. In this situation the user should contact the archive directly.)
- way of delivery (e.g. by email, on CD-ROM/floppy, digital print, photographic reproduction.

The system enables the insertion of information about the customer, but at least
- Organisation
- Name
- Note on privacy of customer.
- Address of delivery

If the end user wants to order another photograph during the session the customer information can be retrieved from cache memory. So the end user does not have to insert the information again.
The system will take the order and redirect it to the archive.

4.1.3 The archive employee

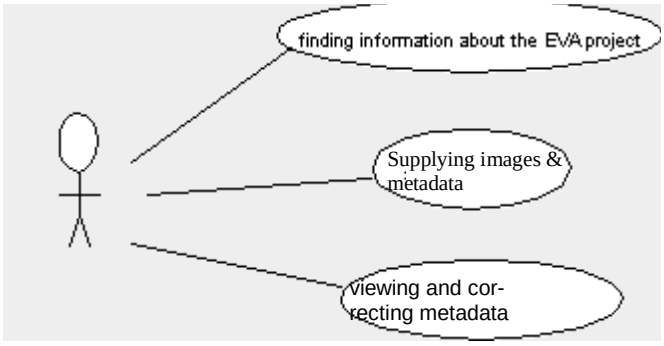


Figure 3: the use case diagram of the archive employee

Finding information about EVA
The employee of an archive with an historical photograph collection wishes to be informed about the EVA project. With this information he can decide to join the organisation. The conditions to join are mentioned.

There will be two such informational needs in an archive. First, archives will be interested in information about the organisational background of the project and the digitisation of photographic material. That way they can join the EVA organisation.

The workpackages of the EVA project should be made accessible through the system.

Second, archives will want to learn about new developments (important updates, the exploitation model, changes in copyright and such).

A newsletter will be published on a regular basis and made available through the EVA system.

Supplying images and metadata
An archive employee will be responsible for adding and updating this information.

The information provided by the archive to the system must comply with the specifications, described separately in Appendix A. These will result in “EVA guidelines”).

Viewing and correcting the metadata
Archives have editorial control over metadata and images. A service level agreement will give details on the upgrade policy.

It should not be possible to overwrite metadata of another archive in the process of adding or updating.
Updates or additions should not lead to huge amounts of network traffic. It should not be necessary to upload the entire collection of metadata at each update.
The system must detect and process the new information in updates and additions.

5. Graphical User Interface design

The Graphical User Interface (GUI) of the EVA-system consists of 11 screens. Each screen will be described in this section. The flowchart has been depicted in figure 4.

The introduction screen contains
- a general and short description of the system and its intentions
- the names of participating archives
- the possibility to select a language for the remainder of the session
- links to
  - the search interface (→ 2)
  - more information for end users (→ 9)
  - more information for archive employees (→ 10)

The search interface contains
- 4 textboxes for input on the following items
  - free text search (on the description elements “title”, “description”, “keyword/subject”)
  - photographer (A list of names of photographers can be retrieved)
  - geography (All geographical terms in original language go in this field, city, street, quarter, etc.)
  - time period (there are two possibilities: (1) user can type in an exact date: YYYYMMDD and images that have this date in the description will be retrieved and (2) user can type in begin year and end year of a period. All images that have a date/period description of which at least one element (e.g. the begin year or end year) falls in the period covered will be retrieved.
- buttons for search (→ 3) and reset
- a button for browsing. (this is in fact a query for all images) (→ 3)
- the possibility to select the collection(s) that will be queried
- if the end user wants to use the multi language technology in his search, he can selects this option
- the possibility to select how the result should be presented
  - both text and thumbnails
  - only text
  - only thumbnails

The result screen contains
- the result as specified in the search interface
- by clicking on a result a larger reference image is presented (→ 4)

The reference image screen contains
- the reference image and all available meta data
- a button for ordering a reproduction of the original photograph (→ 5)

The order statement screen contains
- information about the EVA order procedure
- information about the archive specific procedure and prices
- a button for proceeding (→ 6)

The order form screen contains
- a form for supplying name, address etc.. This form is archive specific.
- a button for proceeding and reset (→ 7)

The order confirmation screen contains
- the information supplied by the end user
- the thumbnail or the reference image of the reproduction that is to be obtained and its meta data
- a button with the caption *buy* (→ 8)

The *Thank you!* Screen
- confirmation of the order
- a button to perform a new search (→ 2)

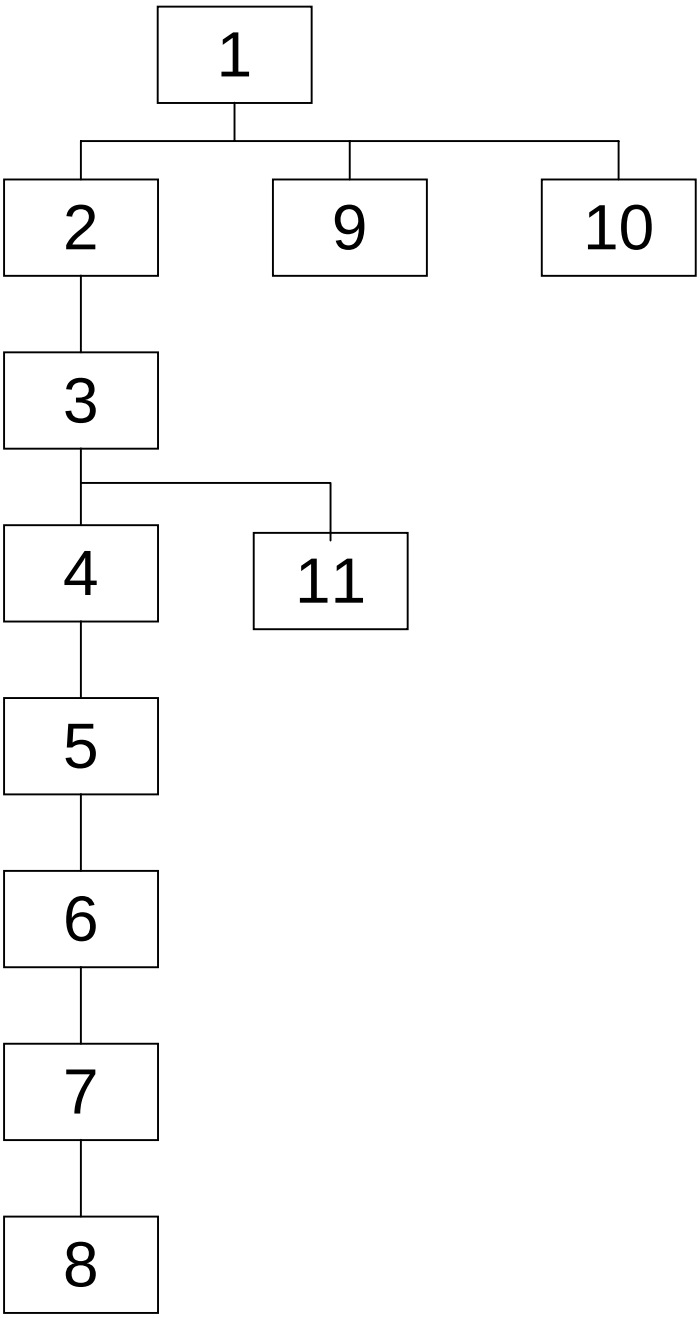
More information for the end users

Information about the EVA project

More information for the archive employees
(link to) information about how to join
(link to) information about how to scan photographic material (workpackages)

Intellectual property statement screen
Archive specific information about the copyrights and the intellectual property of the photograph.

Figure 4: Flowchart of the Graphical User Interface



6. Architecture of the EVA system

According to the functional requirements the EVA system must contain the following information.

Digital images (thumbnails and reference images)
Metadata
Order procedure statement
Intellectual property statement

Images are inserted in a file system. Two directories are distinguished; one for thumbnails and one for reference images.

The metadata can be stored and accessed best with an RDBMS. Insertion of metadata in the RDBMS will be organised by extracting information from an XML-document provided by the archive.

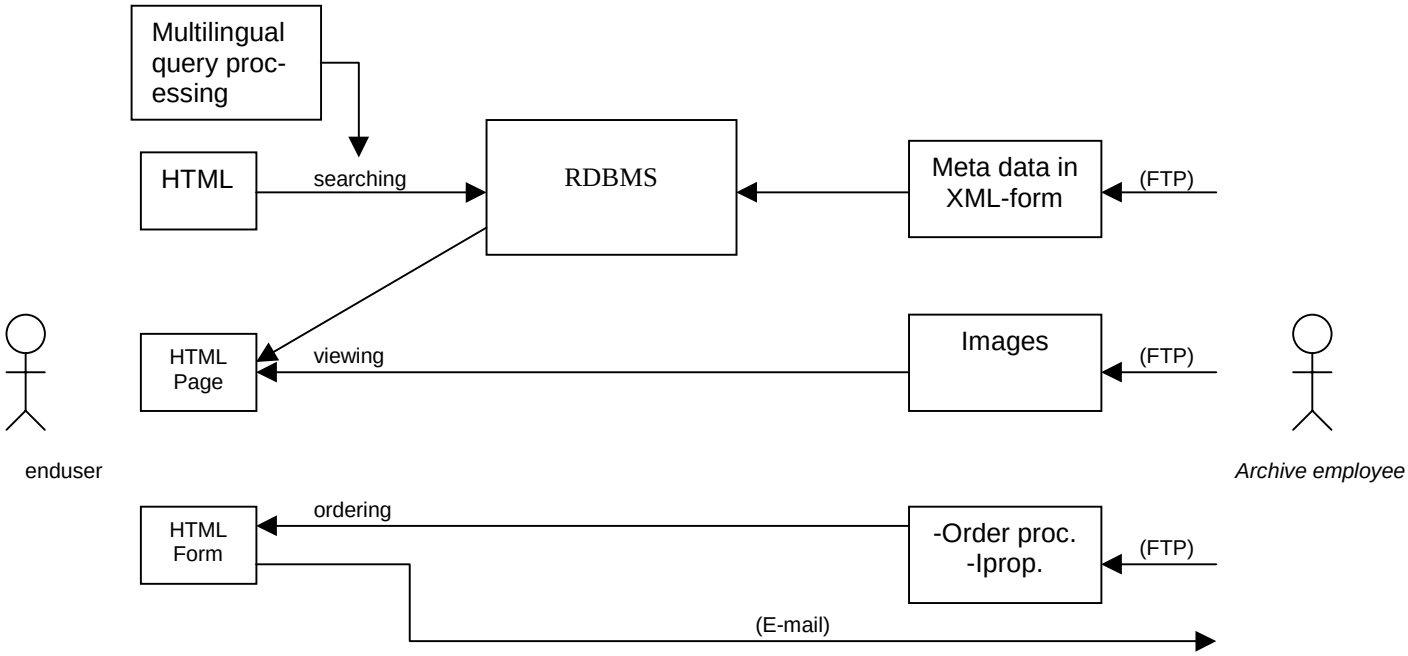
Both the order procedure and the intellectual property statements are describing text, stored best in a file system.

The overview of the architecture is depicted in figure 5.

Multilingual Query Processing

A separate component will be obtained for translating the query into different languages. The translation should be performed between the moment the end-user has sent the query to the system and the actual processing of the query in the RDBMS. The construction is described in appendix D (cf. 'searching' in figure 5).

Figure 5: Global architecture of the EVA system



7. Appendices

Appendix A: Data to be provided by local archives
Appendix B: XML-DTD for metadata
Appendix C: Database structure EVA - system
Appendix D: Multilingual Query Processing
Appendix E: Motivation of interpretation of Dublin Core EVA-system

7.1 Appendix A: Data to be provided by local archives

7.1.1 Images and descriptions

Data to be provided by local archives
The following data should be provided to the EVA-system by the local archives:
Descriptive information of each scene.
Digital images of photographs. Next to the description of an image the archive has to provide the EVA-system with two digital images of each scene: a thumbnail image and a reference image.
Information on the procedures to order an image. Each local archive can have a specific ordering method. It is possible that the buyer receives an invoice together with the ordered print. It is also possible that first the buyer has to pay and that the archive will process the order once the money is received.
Information on the intellectual property rights of the images.

Director and File names
The XML files containing the metadata should be of the form AAANNNNN.XML, where AAA is a unique acronym for a local archive (e.g. LMA / SAA) and NNNNN a sequential unique number.
Thumbnail images: \thumbnail\AAANNNNN.EXT, where AAA is the archive acronym, NNNNN a unique sequential number and EXT the image file extension. The ID number and the image number are related to each other. Related files have the same name. The directory and extension are different.
Reference image: \reference\AAANNNNN.EXT, where AAA is the archive acronym, NNNNN a unique sequential number and EXT the image file extension.
Order procedure statement: AAAorder.txt, where AAA is the archive acronym.
Intellectual property statement: AAAiprop.txt, where AAA is the archive acronym.

Descriptive information
Both LMA and SAA have local information systems to describe its photographs. For the EVA project specific databases are set up e.g. to document the technical characteristics of the digitisation process. The centralised EVA-system requires standardisation of data that is created locally. This means that the local data has to be "mapped" to the EVA-format.

What is XML?
XML stands for 'Extensible Markup Language". An XML document is both human and machine readable. This means you do not need a specific application to get access to the data. Specific applications are available to make editing, indexing, etc. possible. An XML document contains special markup, called 'tags', which usually enclose identifiable parts of the document. Such a part is called an 'element'. There is no pre-defined list of elements. However, there is an optional mechanism for specifying the elements allowed in a specific class of documents. A DTD ('document type definition') contains the elements that are allowed that in a particular type of document. For the EVA-system the EVA.DTD will contain elements that are part of an "EVO" (= related content, images and photographs).

Image specifications

1 Local image format
Depending on local infrastructure and available resources digital masters are created.

2 Reference image (extracted from digital master)
Number of horizontal pixels: determined by aspect ratio
Number of vertical pixels: 300
Number of greylevels: 256 (8 bits per pixel)
Format: jpeg

File name: reference\AAANNNNN.jpg where AAA is the archive acronym, NNNNN a unique sequential number and jpg the image file extension

3 Thumbnail image (extracted from reference image)
Number of horizontal pixels: determined by aspect ratio
Number of vertical pixels: 50
Number of greylevels: 256 (8 bits per pixel)
Format: gif
File name: thumbnail\AAANNNNN.gif, where AAA is the archive acronym, NNNNN a unique sequential number and gif the image file extension.

Order procedure statement

"LMAorder.txt"
LMA requires the following data in order to process orders from EVA:
{a unique LMA/EVA order number}

Name
Address
Postcode
Contact Telephone number
Contact Fax number

Types of copies available:
scanned image on normal paper
scanned image on photographic paper
Cd rom - uncompressed image
floppy disk - compressed image on jpeg file (no liability for corruption during compression/decompression)
new photographic print

Payment terms:
Payment in sterling
Payment in advance
Payment includes VAT
Payment of postage and packaging extra (a minimum of £2)

Rush Fees:
Orders can be processed quicker on payment of a rush fee of £18. It is advisable to contact LMA direct by telephone or email to arrange details of this.

Turnaround time:
Orders are normally processed on a first come first served basis. Items are dispatched from LMA within 7 days of receiving payment but users should allow up to 21 days for delivery. No liability for delays caused by delivery services.

Prices
To be set at a date nearer to completion of the project.

7.1.3 Intellectual property statement

Submitted by LMA at may 23th 2000
(1st draft)

LMA and SAA have undertaken all reasonable actions in order not to violate the intellectual ownership of the images accessible in the EVA-system. LMA and SAA own the majority of the photographs included in EVA and the copyright in them. Other photographs have been included with the permission of the copyright owner. [In the case of LMA a few photographs are Crown copyright. Copyright in some photographs has now expired and in others it is believed to have expired or the identity of the photographer is unknown.] If LMA or SAA has inadvertently infringed the copyright of any photographer or their heirs, please contact the appropriate Archive with details of your claim so this can be rectified.

Copyright in all the scanned images visible in EVA belongs to LMA and SAA. The images can be copied for personal use without restriction. For publication or exhibition in any form permission must be obtained from the Archive which holds the original photograph and the appropriate reproduction fees paid and due acknowledgement made.

For details of reproduction fees see

7.2 Appendix B: XML-DTD for metadata

```
<!-- The information stored in this DTD is a subset of a (future) extended
DTD for an “EVA Visual Object” (EVO). The purpose of this subset is to cap-
ture the information needed for photographic resource discovery.

Change History
Description : Name : Date

Creation : Ivo Zandhuis & Michel Koppelaar (NIWI): 20000508
Small adjustments : IZ & MK (NIWI) : 20000510
-->

<!ELEMENT EVOlite (title, description?, photographer?, (date | time-
period)?, geography*, subject*, archive, location*, relation)     >
<!ATTLIST EVOlite
            xml:lang    NMTOKEN           #REQUIRED
            ID          ID                #REQUIRED              >
<!-- xml:lang maps to DC:Language -->
      <!-- The language must be stated at the outset, in element "EVOlite"
      (the document root)and can be overriden by using this same xml:lang
      attribute in any element inside the root element.
      Value of the attribute must comply with ISO639 (recommended by DC)
-->
<!-- The identification attribute must be identical to the filename of this
EVOlite -->

<!ELEMENT title (#PCDATA)                                          >
<!-- Maps to DC:Title -->

<!ELEMENT description (#PCDATA)                                    >
<!-- Maps to DC:Description -->

<!-- Gives a description of the scene of a photograph. Maximum 500 charac-
ters (about 50 words). -->

<!ELEMENT photographer (#PCDATA)                                   >
<!-- Maps to DC:Creator -->

<!ELEMENT date (year, month?, day?)                                >
<!ELEMENT year (#PCDATA)                                           >
<!ELEMENT month (#PCDATA)                                          >
<!ELEMENT day (#PCDATA)                                            >
<!ELEMENT timeperiod (beginyear, endyear)                          >
<!ELEMENT beginyear (#PCDATA)                                      >
<!ELEMENT endyear (#PCDATA)                                        >
<!-- Maps to DC:Date -->
<!-- Content of element “year”, “beginyear” and “endyear” must
            have the following format: YYYY -->
<!-- Content of element “month” must have the following format:
            MM -->
<!-- Content of element “day” must have the following format: DD -->
<!-- This complies with ISO8601 (recommended by DC) -->

<!ELEMENT geography (#PCDATA)                                      >
<!-- Maps to DC:Coverage -->
<!-- Use a separate tag for every geographical element: e.g.
<geography>Kerkstraat</geography>
<geography>Antwerpen</geography>
in stead of
```

```
<geography>Kerkstraat, Antwerpen</geography>
-->

<!ELEMENT subject (#PCDATA)                                          >
<!-- Maps to DC:Subject -->
<!-- See the remarks about "geography" -->

<!ELEMENT archive (#PCDATA)                                          >
<!-- Maps to DC:Publisher -->
<!-- Name of the participating archive that inserted this EVO -->
<!ATTLIST archive abbrev CDATA #REQUIRED                             >
<!-- Abbreviation of the name of the archive. Unique within the EVA system.
-->

<!ELEMENT location (#PCDATA)                                         >
<!-- Maps on DC:Identifier -->
<!-- path and filename on the EVA-server of the thumbnail or refimg  -->
<!ATTLIST location type (thumbnail | refimg) #REQUIRED>
<!-- Attributevalue "type" indicates whether the refered image is a thumb-
nail or a refimg -->

<!ELEMENT relation (#PCDATA)                                         >
<!-- Maps to DC:Relation -->
<!-- Archival code indicating the location of a photograph in the physical
archive-->
```

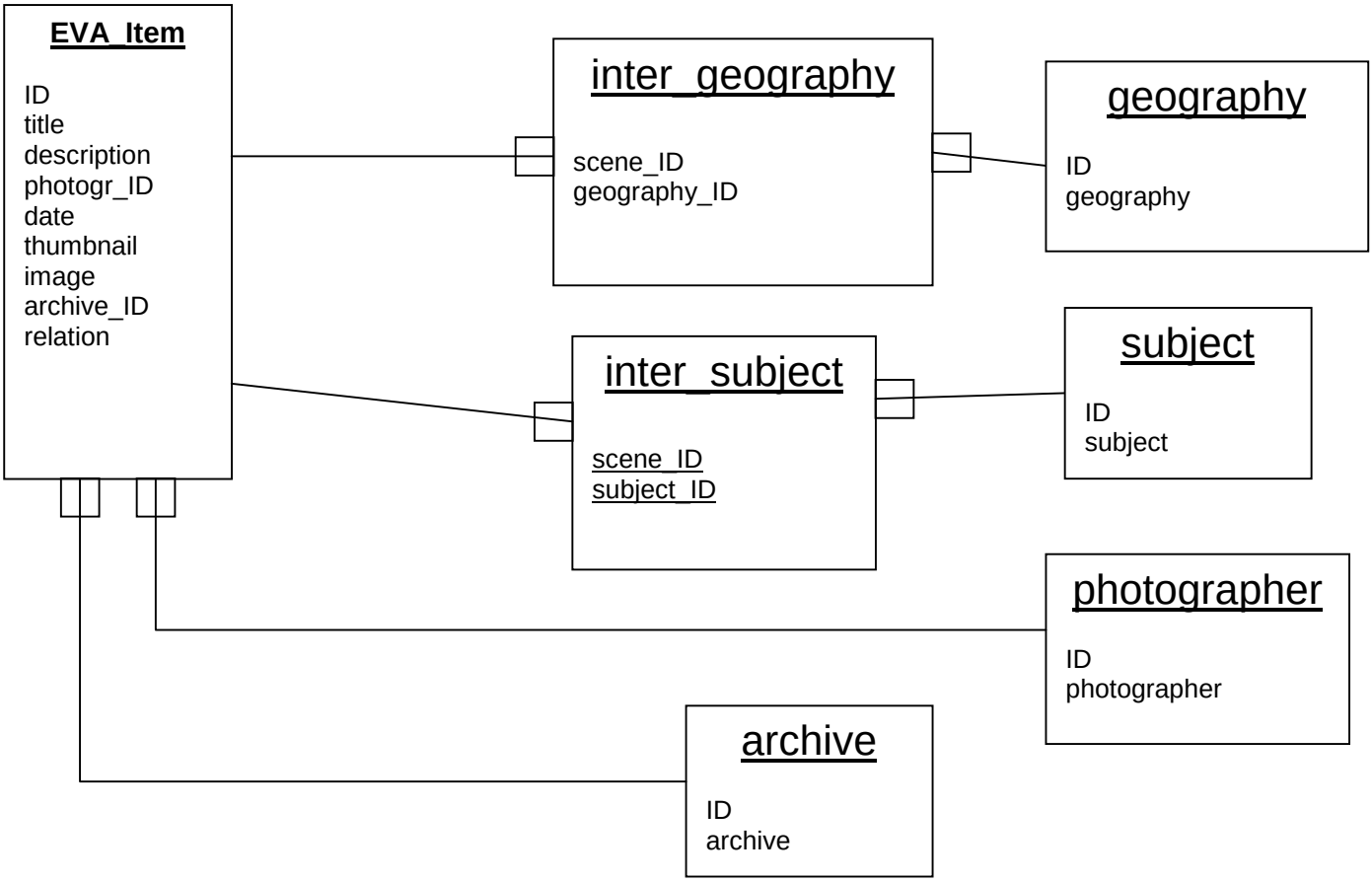
Example XML-documents

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v2.5  - http://www.xmlspy.com -->
<!DOCTYPE EVOlite SYSTEM "EVOlite.dtd">
<EVOlite xml:lang="nl" ID="SAA0001">
       <title>Havenbeeld</title>
       <date>
              <year>1900</year>
       </date>
       <subject>Goederenbehandeling</subject>
       <subject>Granen</subject>
       <archive abbrev="SAA">Stadsarchief Antwerpen</archive>
       <location type="refimg">reference/SAA0001.jpg</location>
       <location type="thumbnail">thumbnails/SAA0001.jpg</location>
       <relation>34031</relation>
</EVOlite>

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE EVOlite SYSTEM "EVOlite.dtd">
<EVOlite xml:lang="en" ID="LMAL04003">
       <title>Gymnasium in new school building</title>
       <photographer>unknown</photographer>
       <timeperiod>
              <beginyear>1930</beginyear>
              <endyear>1940</endyear>
       </timeperiod>
       <geography>Kennington Oval</geography>
       <geography>Lambeth Metb</geography>
       <archive abbrev="LMA">London Metropolitan Archives</archive>
       <location type="refimg">images/LMAL04003.jpg</location>
       <location type="thumbnail">thumbnails/LMAL04003.jpg</location>
       <relation>call: ACC/2692 item: 172/03</relation>
</EVOlite>
```

7.3 Appendix C: Database structure EVA-system

7.3.1 Entity-Relationship diagram for the EVA-system



scene
ID: Unique code to identify a description
title: Brief description of scene visible on image
description:Free text account of scene visible on image
photogr_ID: Foreign key
date: Date and periods relating to scene visible on image- this should be split into different database fields: year, month, day (ISO 8601) and there should be two fields for specifying the first and last year of a time period (the exact date may be unknown)
archive_ID: Foreign key
thumbnail: file name of thumbnail
image: file name of reference image
relation: Inventory number of the original photograph and/or digital master. Also relation with other information system (e.g. EUAN) is possible.

inter_subject
This is just an intermediary entity.

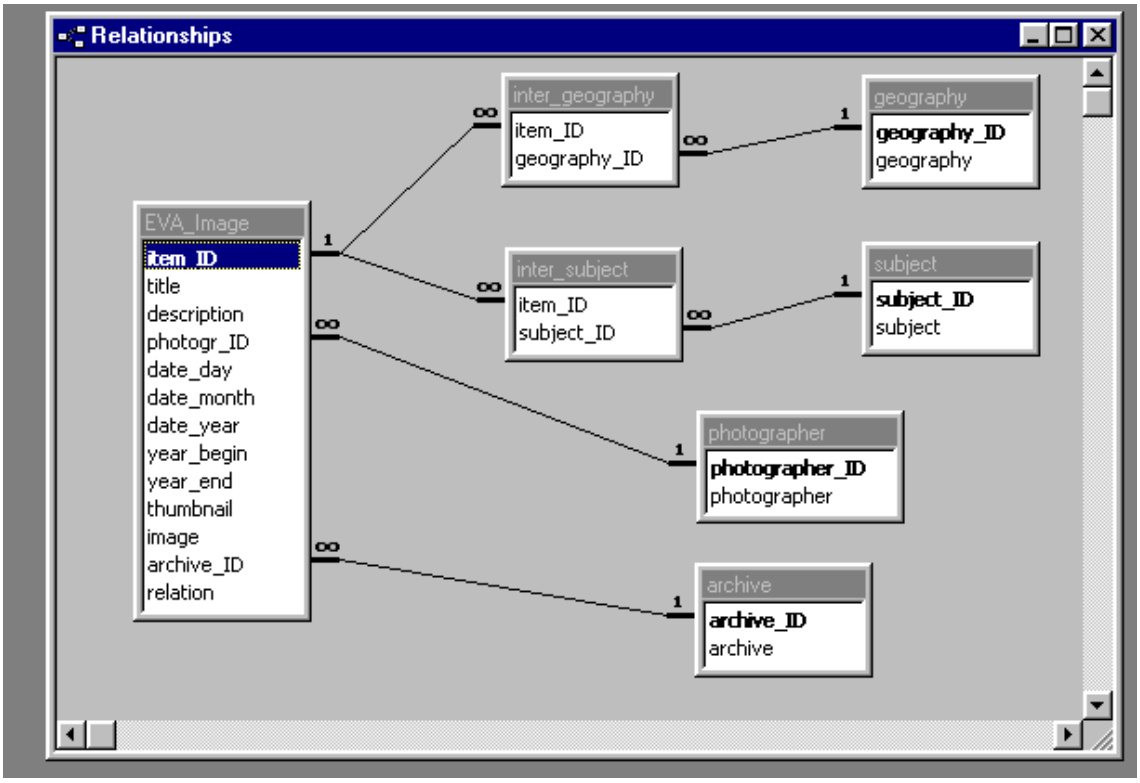
inter_geography
Another intermediary entity.

geography
ID: Primary key
geography: Geographical identifier (street etc.)
archive
ID:Primary key
archive: Name of creator of description

subject
ID: Primary key
subject: Keywords and subject terms

photographer
ID: Primary key
photographer: Name of photographer

Implementation of Entity-Relationship diagram



7.3.3 Mapping database structure with XML-elements of EVOlite.dtd

Name of table	Field name	Mapping with EVOlite element
EVA_image	Title	<title>
"	Description	<description>
"	Date_day	<date(day)>
"	Date_month	<date(month?)>
"	Date_year	<date(year?)>
"	Year_begin	<timeperiod(beginyear)>
"	Year_end	<timeperiod(endyear)>
"	Thumbnail	<location(thumbnail)>
"	Image	<location(reimg)>
"	Relation	<relation>
Geopgraphy	Geography	<geography>
Subject	Subject	<subject>
Photographer	Photographer	<photographer>
Archive	Archive	<archive>

7.4 Appendix D: Multilingual Query Processing

(contribution by Sail Labs, Germany)

7.4.1 Motivation for the actual concept

Multilinguality in EVA
One of the intention of EVA is to let European citizens participate in the cultural heritage of other countries, multilinguality being one of the obstacles to overcome.
Requirements of **multilinguality**, however, need to be analysed in detail:

Users should be able to communicate with the system in their native language. This is a requirements to **user interfaces and communication**. EVA will address this by providing access in at least five European languages, to demonstrate the possibility to extend to even more. The interfaces (dialogues, menus etc.) will be available in the following languages: **English to German, French, Spanish, Italian and Dutch.**

In addition, the system backend should understand **queries**, provide texts etc. This requirement consists of two parts:

We must support the **systems' knowledge base**, which currently is in **English and Dutch**. So users must be able to access English and Dutch resources, in their own language.
While Sail Labs has developed general purpose Machine Translation systems from and into English, there is no similar possibility for Dutch. Developing MT systems for Dutch is beyond the scope of the project, both in terms of duration and funding. Instead, we develop linguistic components to do at least key term translation, to convey the basic content of the texts, and specialised translation tools for term translation. This will be available. We take a third language **(German)**, to demonstrate access to English and Dutch , and to demonstrate in such a way the multilingual capabilities of the EVA system.

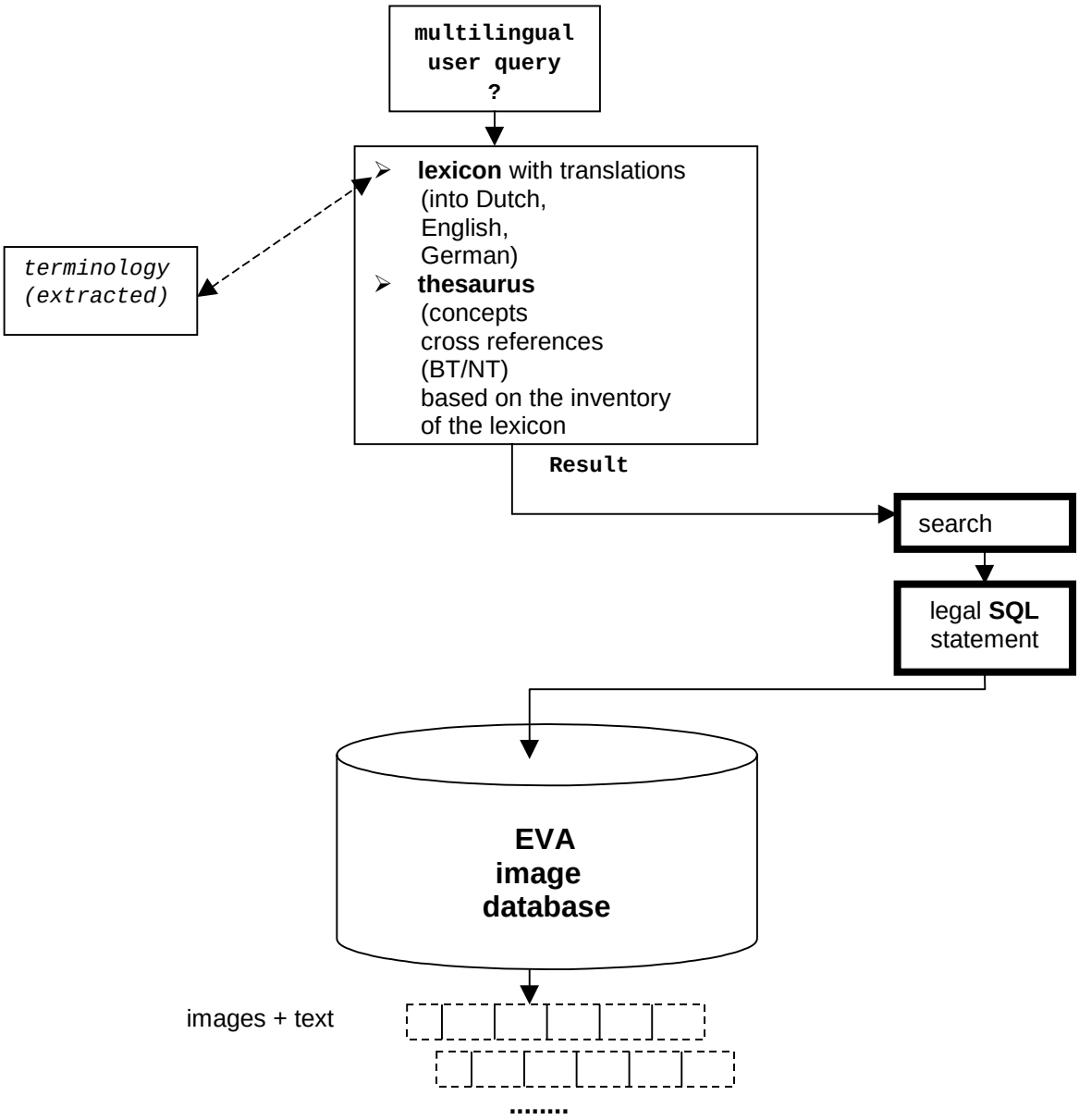
In the framework of the (relatively restricted from the viewpoint of time and funding) EVA project three languages will show the general viability of the multilingual solution. Further languages could be added within further projects.

One of the crucial elements for each multilingual project is the availability of **special domain terminology**. This is an issue which is not a priori supported by general purpose systems, and needs to be taken care of by domain experts. For the EVA domain, the domain specific terminology has to be extracted from the archive description texts. The system needs multilingual lexicons and a thesaurus, i.e. a lexicon has to be extracted, translated, the vocabulary has to be harmonised and a thesaurus has to be defined and connected with the lexicon entries. In order to extend the coverage, this resource needs to be translated into other languages, which is possible but may be beyond the scope of EVA. All different tasks are rather labour intensive and need extensive work..

As a result, we will make available query translation and (partial) term translation on the basis of three languages, integrate such components into the system, and demonstrate their usability in the context of English – Dutch – German translations. If this is successful then EVA as a research project has met its objectives, and extensions to other languages can follow in post-EVA project phases.

7.4.2 Concept for the Query Translation and Expansion

The user can input his query in several languages. The EVA prototype will process queries in Dutch, German and English. The query will be matched by the query translation tool with term translations of the processed languages and with a prototypical set of concept relations which are defined by a prototypical thesaurus.
Thus, the result of the query translation tool, the extracted search string, refers as well to terms of the processed languages as to terms which are connected by related concepts ('thesaurus'); for a first overview see the scheme below:



Functional description :
Task of the QTE (Query Translation and Expansion)

The query translation takes a natural language item and translates it into target languages. It uses a keyword translation approach. It asks the users for the subject area information to disambiguate possible 1:many translations (e.g. 'bank'). It also expands target language synonyms if available. In case a term has no translation this will be reported as well.

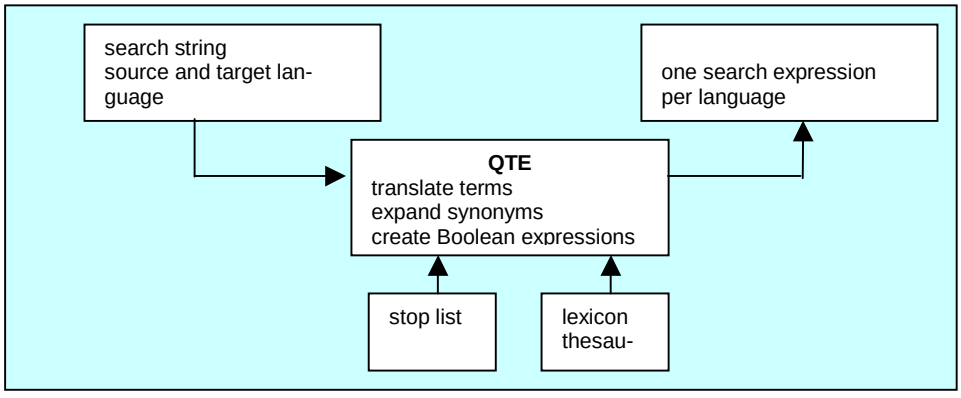
The QTE-Job needs the following input:
a natural language query string, containing the query. Can be one word, or a series of words (a phrase, a complete sentence, etc.)
a list of target languages. This can be ANY, one, or several. In case it is ANY we generate all target languages we can produce. The source language is already known and has to be transferred as parameter.
a list of subject areas for disambiguation. This is a selection from a general list of subject areas. Max. three subject areas are possible.

The QTE processes this input as follows:
it searches for potential translations of the input terms (this can be multiwords). In searching, it applies the subject areas as selection criteria.
it removes all stop words (functional words)
if any content word is left it is converted into a literal, to signal to the user that no translation could be found.
it creates AND links between the remaining parts of the query, as well as between multiword parts
for each language, it looks up all remaining non-literal words to check for synonyms, and ORs the synonyms into the search query.
in case of German, it calls a decomposer to identify relevant compound parts (later version)

The QTE returns an array of records in which a Boolean query expression for each language is available. So the output is
a list of query expressions, one line per language

This output will be converted into a legal SQL statement for searching within the image database.

The input output, and the resources of the Query Translation and Expansion is described in the following diagram:



It would also be possible to put the search directly after the QTE, without user interaction.

Implementation Details

The Query Expansion and Translation (*QTE*) takes a query formulated in natural language and creates Boolean expressions in the original source language and different target languages. There is a COM interface, which makes it easy to use this module from Visual Basic or any scripting language supporting COM objects (like VBScript).

The main object is **QTE** with the program ID "Sail Labs.QTE". It has the following properties:

String
queryExpression

Set and get the expression to be analysed.

String
sourceLanguage

Set and get the source Language of the queryExpression.

By calling the following methods once or several times the client can add one or more subject areas (relevant for translation) and target languages.

addTargetLanguage(String language)

addSubjectArea(String newVal)

To retrieve the analysed and translated expressions the client has to call the method

result

which returns a collection, where each item is a COM Object of type **QTEResult** with the two properties
String
expression

and

String
language

Example:

```
Dim job, jobList
Set jobList = master.jobList(2, 0, "QTE")
Dim Language, Expression As String

For Each job In jobList
   Dim query, queryList
   queryList = job.QTEResult
   For Each query In queryList
      Language = query.Language
      Expression = query.Expression
          ' do something with the strings…
   Next query
Next job
```

Here is an example of how to use the COM Interface in VBScript:

```
Dim qteObj
Set qteObj = CreateObject("SailLabs.QTE")        'creating the COM object

qteObj. queryExpression = "Show me all images of religious buildings in London"
qteObj.sourceLanguage = "English"
qteObj.addSubjectArea "architecture"
qteObj.addTargetLanguage " Dutch "
qteObj.addTargetLanguage "German"

'now retrieve the result:
Dim qteResult
Dim language, expression As String
For Each qteResult In qteObj.result
        language  = qteResult.language
        expression = qteResult.expression
        ' do something with language, expression …
Next


' the result would be something like:

'language      expression
'English       (church OR cathedral OR chapel) AND London
'Dutch         (kerk OR kathedraal OR kapel) AND London
'German                (Kirche OR Kathedrale OR Kapelle) AND London
```

This search expression will be converted into a legal SQL expression which can execute a database search within the image database.

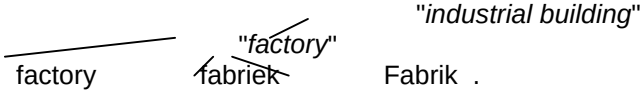
7.4.3 Multilingual Query Processing in EVA: Components and workflow, new Languages

Components

The search for images within the image databases performes an evaluation of the text which describes the images. As the user requires hits from databases of different countries in different languages, the search has to touch heterolingual equivalents of the input search string as well. In order to perform this operation the query should be multilingual. For this purpose EVA uses a multilingual query translation and expansion tool. This Query Processing tool (QTE) applied in the project consists of the following components:
lexicon
thesaurus
lemmatizer
query processor.

Lexicon
The lexicon is built on the basis of lexical material which is used in the image databases of the participating archives for titles, keywords and descriptions of the stored images. The entries of the lexicon are lexical items extracted from the texts and evaluated by specialists. The content of a lexicon depends on the dimension and the content of the database. The lexicon contains non-inflected canonical word forms with a certain number of linguistic features.
Each database and query language needs a lexicon with the lexical material used in the descriptions texts. The lexicons derived and extracted from the texts, have to correspond to the domain which is represented by the archives in order to get satisfying search hits.
If archives with different description languages are included into one and the same query process, satisfying results can only be obtained if the lexicon is supplemented by translations of terms of language 1 into terms of language 2 until language n. In order to make easier this operation the terms of texts descriptions in different languages can be extracted from the texts of the different languages - if a term extractor is available. But the term lists should be harmonised with the term inventory of the original language. In such a way bi- and multilingual lexicons with 1 to 1 (to 1 ...) equivalents can be created.

Thesaurus
In order to achieve more extended hits and improved results of the database search, the search string has to be connected with a network of related expressions. The network is represented in the form of a thesaurus which defines broader and narrower relations of abstract concepts within a structured hierarchy. The nodes of the thesaurus are conceptual units and should be language independent, i.e. one and the same concept is related with terms of different languages. They have a link to all corresponding multilingual terms (= entries) of the lexicon. The representation language of the thesaurus for concepts is English, so that an English concept named 'factory' with broader term links to another English concept 'industrial building' has term links to the English, Dutch and German terms 'factory', 'fabriek', and 'Fabrik':



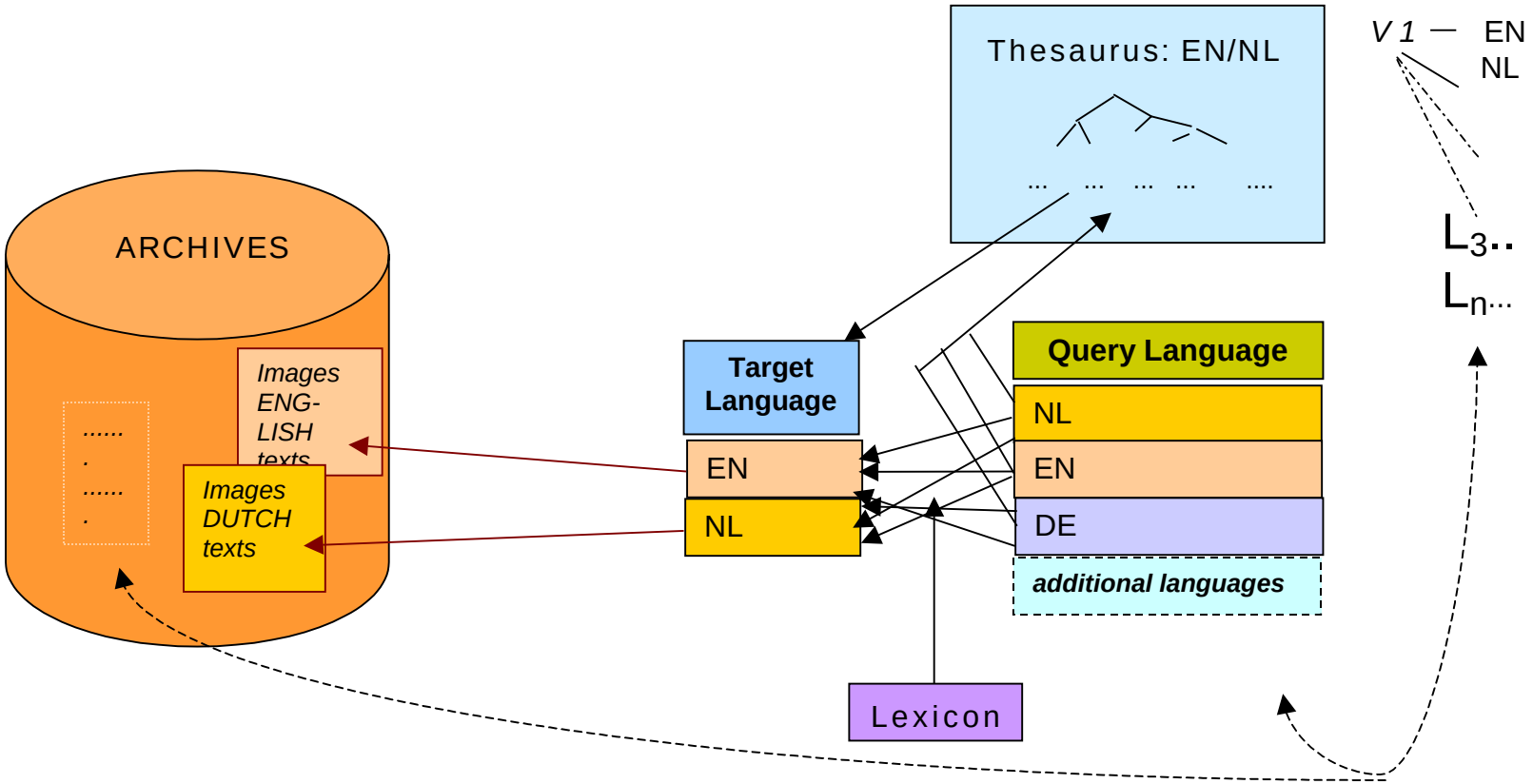
The thesaurus has to be defined, created and maintained by domain specialists.

Lemmatizer
The lemmatizer is an autonomous tool which interacts with the query processor. The lemmatizer reduces inflected word/text forms to a canonical word form and distributes linguistic features to certain entries. As inflected text word forms are not stored in the lexicon lemmatizers are mandatory components of the query processing tool for multilingual searches.

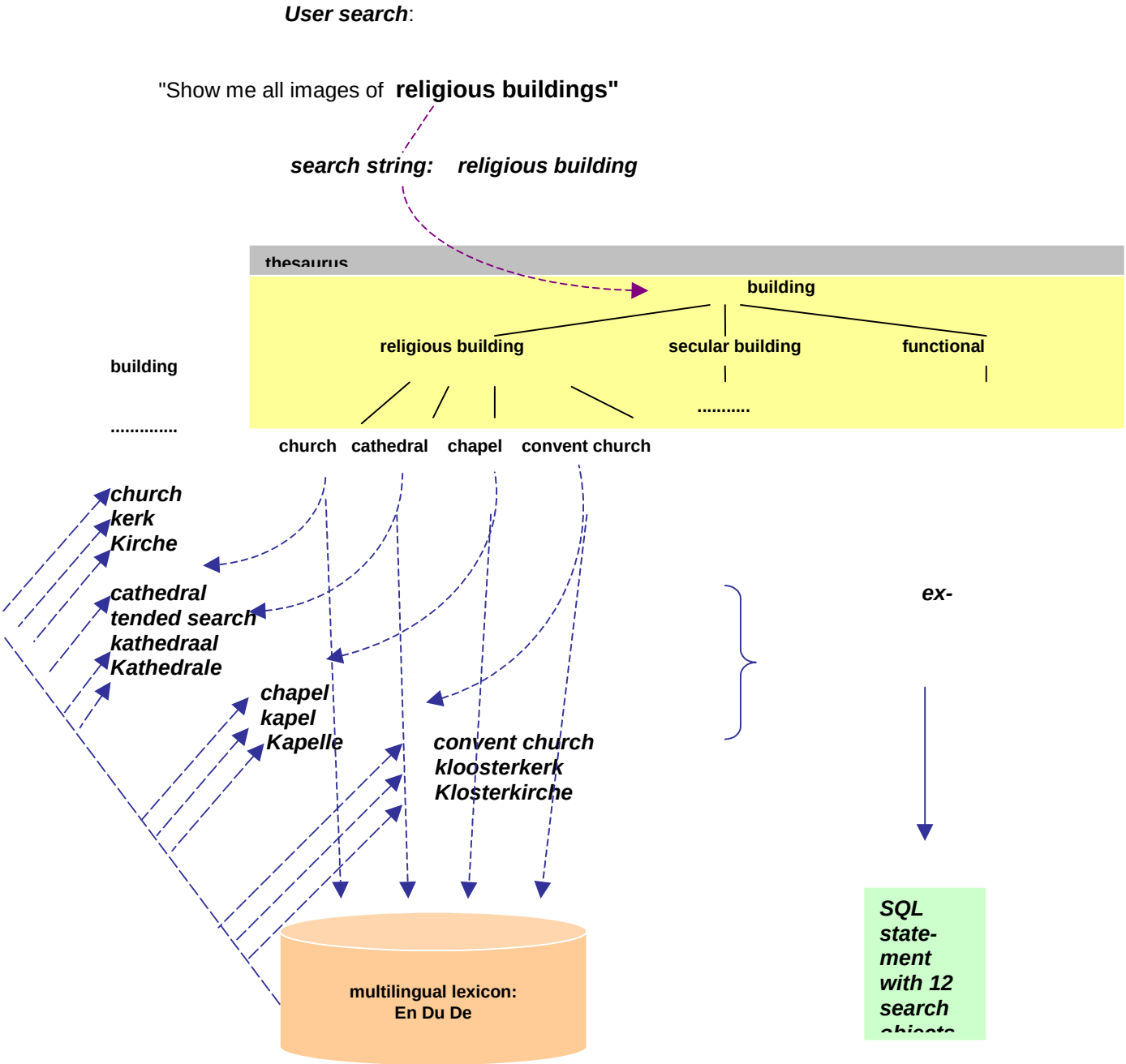
Query processor
The query processor represents the software frame which triggers the different components and calculates the result search string which is transferred to the image database. The final result is delivered in the form of a valid SQL statement according to the structure of the image database.

The basic architecture is described in figure 1:

Figure 1:


Thesaurus: EN/NL
... ...
V 1 — EN
NL
L3..
Ln...
ARCHIVES
Images ENG-LISH texts
.....
.
.....
.
Images DUTCH texts
Target Language
EN
NL
Query Language
NL
EN
DE
additional languages
Lexicon

Example:
Using a lexicon with an integrated thesaurus, a search string 'religious building' would access not only the keyword 'religious building' and its pre-translated heterolingual equivalents but also the related expressions which had been defined within the thesaurus structure, cf. the following scheme:



Interaction:
Thesaurus and lexicon are modelled in a database which interacts with the query process component. The underlying data model consists basically of three main tables: terms, concepts (i.e. units of the thesaurus) , and links. These tables are supported by definition and administration tables.
Concepts as units of the thesaurus are abstract elements. They have a “name”, a category , an optional definition, and also some subject area code (for disambiguation e.g. bank).
There are links between words of difference languages (i.e. translations, like en, du, de), and links between terms/words and concepts. (cf. above: broader term, narrower term...).

Import:
The lexicon needs to be connected with supplementary lexical resources for updates or for adding further languages. For this purpose a special lexicon interchange format is used: OLIF (Open Lexicon Interchange Format).. The first version of OLIF was defined in the framework of the EC OTELO project. An OLIF standard will be defined in the framework of the OLIF consortium (members are e.g. Microsoft, SAP, Logos Sail Labs, Lotus, IBM).
The lexicon communicates with external data sources by reading entries in the OLIF format and distributes entry features and values from an OLIF input file into the respective internal tables (= *import functionality*).

OLIF is based on a XML/SGML type notation. Each OLIF exchange file has a XML/SGML header containing basically global definitions of features and values used in the OLIF document type definition and a body containing the entries. Basic items are features and values. Features are defined as elements, with a start and an end item. The value is written between these two items:
e.g. ' <POS>noun</POS> '
= indicating that the feature *part of speech (POS)* has the value *noun*.

Some of the features in every entry are obligatory, as they are necessary to identify the entries. They must be available in the lexicon system which is to be interrelated by OLIF. The following features uniquely define an entry:

canonical form - language - part of speech - subject area

They are used not just to identify an entry but also to address the target entries for the link features.
To describe the *meaning* of the target entry OLIF proposes to use *subject areas* which de facto *are* a semantic classification of entries; cf. the example:

```
<Entry>
        <MONO>                          ; this describes the source of the
link
                <CAN>dog</CAN>
                <CAT>noun</CAT>
                <LG>En</En>
                <SA>GV</SA>
        </MONO>
        <XFR>
                <CAN>chien</CAN>        ; this describes the target of the link
                <CAT>noun</CAT>
                <LG>Fr</LG>
                <SA>GV</SA>
        </XFR>
</Entry>
```

Adding new languages:

In the framework of the EVA project the query can be performed in Dutch, English and German. The languages of the archives are Dutch (SAA) and English (LMA). A query in German will find images of Dutch and English image pages in the database, an English query Dutch and English pages etc.

New languages can be integrated as

additional archives in other languages as English, Dutch and German,

- e.g. French if a French archive will be integrated with description texts in French

additional query languages

e.g. an extension of the number of the query languages will also permit to find respective images if the user input a query string in French, Italian, Spanish or other languages.

In order to integrate further languages the following work packages have to be done:

create new lexicon for the new language on the basis of the text descriptions of the new archives

evaluate the new lexical material

harmonise the new material with the already existing lexicon entries (content, translations etc.)

create a new thesaurus or adept the new lexicon to the existing thesaurus

adept a lemmatizer of the new language to the query process.

The manpower needed for a lexicon with approximately 10 000 entries of a new language is approximately 6 MM.
To develop a new lemmatizer needs at least 10 MM depending on the language.

The creation of a thesaurus or the adaptation to an existing thesaurus can only be estimated after the experiences with the thesaurus of the languages in the current EVA project.

Updates and upgrades:
Updates and upgrades will be performed using the OLIF import interface of the lexicon database.
The following steps are necessary;

extraction of new lexical material
evaluation of new lexical material
translation of new lexical material
harmoisation of new lexical material
connection to the exixting thesaurus.

The manpower needed for these operations is identical with the manpower needed for new language lexicons:

The same adaptation and integration steps have to be carried out. The manpower for the adaptation to the thesaurus can be estimated only after experiences with the existing thesaurus.

7.5 Appendix E: Motivation of interpretation of Dublin Core by EVA-system

The Dublin Core (DC) is a metadata element set intended to facilitate discovery of electronic resources. A wide range of organisations, like museums, archives and libraries uses the Dublin Core. The Dublin Core consists of 15 elements. The Dublin Core website provides detailed information on the use and status. The address of the website is: http://purl.org/dc/. In the following table the 15 DC elements are grouped according to content, intellectual property and instantiation. In case the EVA-system uses a DC element a short description in italic is given.

CONTENT	**INTELLECTUAL PROPERTY**	**INSTANTIATION**
DC.Coverage *EVA: Geography*	DC.Contributor	DC.Date *EVA: Date photograph is taken*
DC.Description *EVA: Description*	DC.Creator *EVA: Photographer*	DC.Format
DC.Type	DC.Publisher *EVA: Archive*	DC.Identifier *EVA: ID of photograph – digital image*
DC: Relation *EVA: Relation with image / photograph*	DC.Rights	DC.Language *EVA: Language used in title / description*
DC.Source		
DC.Subject *EVA: Subject / Keyword*		
DC.Title *EVA: Title*		

Figure 1. 15 Dublin Core elements grouped by intended use. In Italic the corresponding EVA-system element.

The EVA-system uses 10 of the 15 Dublin Core elements: Title, Description, Creator, Date, Relation, Coverage, Language, Subject, Identifier, Publisher. The following DC-elements are not used: Contributor, Source, Type, Format and Rights Management. The reason for this limited use of the DC elements is that the current and future content providers of the EVA-system have a low threshold to join the EVA-system, whereas a much more extended local "back-office" at least supports all DC elements. It is the intention of the EVA-project to develop an EVO.DTD that covers all attributes of all objects that are relevant for the dissemination of historical photographic collections. The development of the EVO.DTD is beyond the scope of the EVA-system.
The exclusion of the DC element Rights Management is motivated as follows: the selection of the photographs in the EVA-system is based on a clear heterogeneous copyright situation. A general "copyright statement per archive" (See appendix 7.1) will cover all photographs.

The EVA-system is based on the EVOLite.DTD (see appendix 7.2). This DTD contains elements that can be mapped with Dublin Core. The following DC-elements are used in the EVA-system.

1 Name: Title
Identifier: Title
Definition: A name given to the resource.
Comment: Typically, a Title will be a name by which the resource is formally known.
Mapping with EVA-system element: Title.

2 Name: Creator
Identifier: Creator
Definition: An entity primarily responsible for making the content of the resource.
Comment: Examples of a Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.
Mapping with EVA-system element: Photographer.

3 Name: Subject
Name: Subject and Keywords
Identifier: Subject
Definition: The topic of the content of the resource.
Comment: Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
Mapping with EVA-system element: Subject.

4 Name: Description
Identifier: Description
Definition: An account of the content of the resource.
Comment: Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.
Mapping with EVA-system element: Description.

5 Name: Publisher
Identifier: Publisher
Definition: An entity responsible for making the resource available
Comment: Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Mapping with EVA-system element: Archive.

6 Element: Date
Name: Date
Identifier: Date
Definition: A date associated with an event in the life cycle of the resource.
Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and follows the YYYY-MM-DD format.
Mapping with EVA-system element: Date.

7 Element: Identifier
Name: Resource Identifier
Identifier: Identifier
Definition: An unambiguous reference to the resource within a given context.
Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Mapping with EVA-system element: Thumbnail / Ref. Image (name and directory of Thumbnail and Reference image). The element Identifier is related with the images in the EVA-system.

8 Element: Language
Name: Language
Identifier: Language
Definition: A language of the intellectual content of the resource.
Comment: Recommended best practice for the values of the Language element is defined by RFC 1766 [RFC1766] which includes a two-letter Language Code (taken from the ISO 639 standard [ISO639]), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard [ISO3166]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.
Mapping with EVA-system element: Language

9 Element: Relation
Name: Relation
Identifier: Relation

Definition: A reference to a related resource.
Comment: Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
Mapping with EVA-system element: Relation (archival code indicating the location of a photograph / images in the physical archive. The function of the element Relation is that the archive can locate both the image and the photograph. This element expresses the relation of the EVOLite 'object' with the photograph and digital image.

10 Element: Coverage
Name: Coverage
Identifier: Coverage
Definition: The extent or scope of the content of the resource.
Mapping with EVA-system element: Geography. This element contains all geographical descriptions (street, quarter, city, etc.) relevant for the scene on the image/photograph.