CIDOC 2015 New Delhi

An Introduction to PHAROS: Aggregating Free Access to 31 Million Digitized Images and Counting . . .

Part 1: Project Overview – David Farneth, Getty Research Institute (dfarneth@getty.edu)

Part 2: Data and Technology Issues – Regine Stein, Bildarchiv Foto Marburg (with contributions from Francesca Mambelli and Francesca Tomasi, Fondazione Federico Zeri) (r.stein@fotomarburg.de)

Part 3: Image Scanning at the RKD – Remigius Weda, Netherlands Institute for Art History (weda@rkd.nl)

Part 1: Project Overview – David Farneth, Getty Research Institute

[slide 1]
Thank you to the program committee for giving us this opportunity to introduce the PHAROS project to the CIDOC community.

[slide 2]
Our presentation today is in three short parts: I will give an overview of the PHAROS[1] project, Regine Stein from Bildarchiv Foto Marburg will talk about some of the technical considerations of the project, and Remigius Weda, from the Netherlands Institute for Art History, will describe the rapid-capture scanning project they are undertaking to digitize their photo archive.

[slide 3]

We thank Inge Reist from the Frick Art Reference Library in New York for proposing this session and contributing to it. Inge is the project coordinator and chief spokesperson for PHAROS, and has been the organizing force since the beginning. The project has also benefitted from financial support from the Samuel H. Kress Foundation to support the cost of meetings and consultant services.

[slide 4]

While our project focuses on historical photo archives at the present, we hope to partner with many of your institutions when this project moves from the research and testing phases into full-scale implementation. If we have time at the end, we hope to hear your ideas about how this project might benefit your work.

[slide 5]

What are photo archives?

The fourteen collections in the initial group were founded as long ago as the late 19th century and as recently as the 1980s. They were amassed by institutions or scholars as a tool for studying art and architecture during a period before the development of a robust market for art books, before international travel become affordable to more people, and, indeed, before the internet. They tend to focus on reproductions of works of art as well as historical sites, monuments, and buildings. The archives are usually organized as browsing collections, with the photographs arranged according medium (e.g "paintings"), country of production, and/or artist name. Added to many of the photographs over time are information-rich layers of annotations, made either by the scholar who collected them or subsequent scholars who have studied them. In this regard, photo archives possess a wealth of unpublished information otherwise inaccessible to researchers.

[slide 6]

Collaborative project of 14 institutions

"PHAROS: An International Consortium of Photo Archives" is a collaborative project among fourteen photo archives in North America and Europe established to create a common platform for research on images of works of art in all media, both Western and non-Western, through comprehensive consolidated access to photograph archives. It will be a freely available commons designed to stimulate research in a broad spectrum of fields by linking together for the first time all of the images and their formal documentation with knowledge amassed by scholars over the years.

[slides 7-13]
Initial participants

The initial project participants and the size of their collections:

Bibliotheca Hertziana, Rome (1,065,000)
Bildarchiv Foto Marburg, Germany (2,000,000)
Courtauld Institute of Art, London (4,173,500)
Fondazione Federico Zeri, Bologna (290,000)
Frick Art Reference Library, New York (1,346,000)
Getty Research Institute, Los Angeles (2,086,000)
Villa I Tatti, Florence (239,000)
Institut National d'Histoire de l'Art, Paris (750,000)
Kunsthistorisches Institut, Florence (650,000)
National Gallery of Art, Washington (7,600,000)
Paul Mellon Centre, London (185,000)
Rijksbureau voor Kunsthistorische Documentatie, The Hague (7,000,000)
Warburg Institute, London (3,500,000)
Yale Center for British Art, New Haven (132,000)

[slide 14]
This project really "takes a village" to accomplish, as you can see from this photo of the people who gathered for the first organizational meeting at the Frick in January 2013.

[slide 15]
What are we aggregating?

The platform will be dynamic, growing over time both in number of works of art represented and an ever-expanding group of institutional participants, including – we hope -- those from Asia, Africa, and Latin America. Current members of the consortium are in a position to begin providing access to hundreds of thousands of the more than 31 million images they hold that document works of art in varying states and at different moments in time. A unique feature of the project calls for the scanning and aggregation of all of the annotations made on the photographs by eminent scholars over the last 100+ years, thus providing a pan-cultural view that would be impossible to achieve in the analog world, even as the photographs themselves retain value as historic documents.

In short, duplications of the same image are very important to the project, and we are as interested in aggregating the unique metadata and information on back of the image as we are with the image itself. For this project, the image functions as both image and metadata.

[slide 16]
What are the research benefits?

Consolidated access to tens of millions of images of works of art will be of immeasurable value to scholarship and teaching for a wide range of art-historical research topics including provenance and attribution, conservation research, exhibition research, publication history, the history of photography, the history of art history, and a myriad of other points of information that we can hardly imagine. [slide 17] We also anticipate an exciting array of new research brought about by leveraging linked data, data visualization, computer vision, and big data aggregation and analysis.

The project will also foster a greater understanding of the photograph as physical object and its critical role in the dissemination of knowledge.[2]

[slide 18]

The benefits of the project are numerous and varied, and they will serve a wide range of people interested many aspects of art and art history. In the published paper available online you will wind a longer list of project benefits and research outcomes.

[The following list will not be read at the live presentation.]

- Expose researchers to the vast knowledge held in photograph archives that is currently hidden and under-explored
- Advance the founding missions of the participating institutions by making millions of digitized images, and their metadata, freely available for research
- Facilitate the discoverability of dispersed visual materials on a network level, including rare and unique negatives, x-rays, and prints
- Harness new and emerging technologies, including computer vision, linked data, data visualization, and big data analysis for the benefit of art historical research and the digital humanities
- Enable the discovery of similar, variant, and related works of art, comparing specific content, shapes, composition, color, etc. through the use of computer vision
- Provide a broader and deeper understanding of factual information and art-historical discourse about the work of art as created over time on different continents, for different purposes, and from differing cultural perspectives
- Enable cross institutional search results to bring forth broad and deep comparisons of the physical changes that have impacted a given work of art over time. Such comparisons will encourage assessment from the points of view of conservation, art history, and changing attitudes toward artistic styles in different cultural contexts and historic periods

- Enable study of the photograph itself (photographic method, date taken, by whom, for what purpose, provenance, publication history, etc.), which is the information needed to study the influence of photography on historical narratives, and historiography of art history
- Enable the re-purposing of data for new research applications

[slide 19]
Where are we?

A number of pilot projects have been identified as proofs of concept. In 2014, the Kress Foundation supported a study to test the viability of image recognition technology on a test set of images and metadata related to Italian Anonymous 15th century paintings, drawings, and sculpture. The success of this project then led to the next pilot project to build a new image searchable database and customized interface. This new technology platform will feature image browsing and text searching, image similarity analysis, searching by image, a data upload interface, and a multilingual interface. A fuller description of this platform can be found on at: http://ejohn.org/research/italian-art-database-proposal/

The project is investigating the use of ResearchSpace for the collaborative environment, an open-source platform developed with funding from the Andrew W. Mellon Foundation and administered by the British Museum, and the CIDOC Conceptual Reference Model (CRM) as the knowledge representation for the metadata (http://www.researchspace.org/ and (http://www.cidoc-crm.org/). The project will continue investigating the use of computer vision to aid in searching, aggregation, and metadata analysis and enhancement.

+ + +

Part 2: Data and Technology Issues – Regine Stein, Bildarchiv Foto Marburg (with contributions from Francesca Mambelli and Francesca Tomasi, Fondazione Federico Zeri)

[slide 20]
What are the specific data and technology issues that we are addressing with this project?

[slide 21]
The databases of the institutions participating in PHAROS embody two different functions:
- **Catalog**: The description of objects (photographs) actually owned
- **Repertoire**: The description of objects not owned – the subjects of the photographs. They are probably described in the catalogs of the institutions that preserve the works of art and/or in those of other photo archives that include photographs of the same works.

We will briefly highlight the two most important aspects in documenting photographs: their dual nature as monument versus document, and their dual nature as serial versus unique objects.

[slide 22]
The photograph as monument (or physical object)
It is the cultural heritage object itself. In the collection management system of a museum it will be described as a unique artwork, with the depicted object often only mentioned in the title.

[slide 23]
The photograph as document
In the database of a monument preservation office the photograph typically serves as a visual representation of the real-world object, with no information about the photograph as such.

[slide 24]
Photo archives typically address both the photograph as a physical cultural heritage object AND as a document of the depicted object. Yet the catalogs have different structures. Data about artworks, historical sites, buildings, and about the photographs are more or less separated or integrated. We thus have two models:

- one providing a single entry with sections devoted to the subject and others to photographic object itself
- one providing two different cataloging units, more or less independent of each other.

[slide 25]
In the perspective of sharing data between different institutions, PHAROS has identified CIDOC CRM as the privileged ontology to which to map the databases of the consortium's members. Some institutions have already started the mapping and transformation processes, such as Fototeca Zeri, Bildarchiv Foto Marburg, RKD etc.

[slide 26]
Since photography is a "translation" of an artwork "by means of another art", each image offers a different interpretation of the works. Therefore, a photo cannot be impersonal and objective. Each positive print, although deriving from the same negative, as noted, mounted, classified differently within the various archives, conveys different information. In particular, annotations on the back, classification, documentation attached to photos, data reported in the catalog records, all together transmit all the different attributions that scholars (and institutions that have used and/or stored photographs) have proposed for every single work of art. The integration of data and images will allow reconstructing not only the history of the works of art, but also the history of *connoisseurship* and the history of the history of art.

The example here presented shows the potential for art-historical research that might arise from the sharing of images and information. It refers to photographs and relative entries about a single work of art traced in three different databases of institutions belonging to PHAROS:

[slide 27-30]

Fondazione Federico Zeri, Bologna
Villa I Tatti, Florence
Frick Art Reference Library, New York

[slide 31]
This aggregation of images offers 9 different visual sources for studying the painting, taken in different contexts and documented in different states of preservation.

[slide 32]
The merging of data allows the reconstruction of multiple steps in the artwork's conservative and attributive story. As shown by the example, the databases of the institutions are characterized by richness, analytics, and scientific reliability of metadata about the works of art and, often, about photographic objects.

[slide 33]
In recent times, we are facing a gradual increase of interest by institutions that own the photos towards the photographic object itself, promoted from a mere documentary function and from its total subsidiarity to the work of art depicted. This happens especially in the approach to historical materials (photographs before 1920), items that represent important sources not only for art history but also for the history of photography.
Each photograph, for the mechanisms of production by which it is obtained, is characterized by:
- **Seriality**: A high number of prints can be taken from the same negative; some editorial series are comparable to editions of books etc.
- **Uniqueness**: Even in the case of positives taken from the same negative, prints can be obtained with different techniques, at different times and so on.
The uniqueness of a photo is also due to the use that was made of each item after its production, e.g. for editing, annotation, reuse for publications

[slide 34]
As mentioned earlier, PHAROS has decided to use the CIDOC CRM as the privileged ontology to which to map the databases of the consortium's members. The next step will be a pilot to integrate CRM mappings from some PHAROS institutions into the ResearchSpace

platform using its 3M Mapping Memory Manager, and to analyze data integration results across different institutions.
Also, colleagues from Zeri are already further investigating specific mapping questions such as attributions of the artworks transmitted from the photographs and/or from the structure of the archive.

[slide 35]
Remigius Weda will describe a new project for rapid digitization of the RKD's photo archive, which is one of the largest in the world devoted to art history.

+ + +

Part 3: Image Scanning at the RKD – Reem Weda, Netherlands Institute for Art History

[slide 36]
Thank you.
Let me start by mentioning that what I'm going to say about our mass-scanning project is theoretical at the moment, because the project hasn't started yet. Fortunately we have a plan for it, and our organisation has a strategic goal to work towards a digitally oriented service. So, now that is out of the way, let us continue…

[slide 37]
RKD Netherlands institute for art history started its existence in The Hague in1932. Here you see the *Hofvijver, the pond in front of the parliament buildings in 1*692 in a painting by Gerrit Berckheyde. This is a picture of the square in front of the old parliament buildings today, with the high-rise government ministries behind it.

[slide 38]

Presently located in the Royal Library building complex, we staff about 65 FTE + volunteers and trainees. The RKD supports organisations and individuals in conducting research, organising exhibitions, or writing publications. For more than 80 years now, we have been working together successfully with institutions and individuals both in the Netherlands and abroad. The RKD is a member of the (Dutch) Museums Association, the Dutch Postgraduate School for Art History (OSK) and a founding member of the International Association of Research Institutes in the History of Art (RIHA).

[slide 39]

Physical collections. Vast collection of photos and reproductions of artworks, estimated about 6 million, and only 200,000 are digitised and available online. Library – about 450,000 volumes. Archives – about 1.5 km. (600 archives).

[slide 40]

Researchers, professionals, interested laymen, and students make use of the documentation for art historical research, not in the least for information for provenance of certain works.

[slide 41]

RKD has made the strategic decision to change focus from analog to digital collections. Why? We want to reach a larger audience. The number of physical visitors to the RKD study rooms is only 5,500 visits a year. The number of visitors online is many times this amount, nearly one million, since we launched our new portal for the datasets. We are aiming for a larger public and want to expand our online presence. The first step was to develop a new portal for our digital assets, called RKDexplore.

[slide 42]

There are 7 datasets in RKDexplore, of which the most important are: RKDimages and portraits – 210,000 works of art with descriptions, biographical data on the depicted persons, and with links to our documentation. RKDartists – biographical information about 300,000 artists, collectors, art historians, etc.

PIC: Like Paulus Potter masterpainter who died young, here in a portrait by Bartolomeus van der Helst, another 17th century masterpainter. PIC: RKDartist is our prime digital access point to the physical documentation and library collection.

[slide 43]

Currently we digitize (parts of) the documentation collections with an average of some 17,500 photos every year (in ca. 15,000 records). Our present strategy is to present our users with photographic material in combination with metadata of the highest standards.
PIC: At this rate, if we want to present all of our art documentation in digital form, it will take some 285 years before we're done. So we need another approach!

[slide 44]

The plan that was conceived soon came to be known as 'RKDboxes'. The idea is to have all the boxes and their contents scanned in a "rapid-scanning-street." Metadata will be added to every object from the information on the box labels. In addition, the texts associated with every object will be scanned and OCR-readable, so that searches will cover any typed annotation with the documentation.
PIC. Here we see an example of a scanning street in another Dutch institute, the Naturalis Biodiversity Centre.

[slide 45]

PIC: Naturalis is scanning their massive collection on natural history comprised of a staggering 7 million objects. Not all objects are scanned separately in this process; they also do entire drawers of specimens. They have 10 scanning streets operating simultaneously, and nearly 80 people working on the whole process. This project is huge. The costs are 13 million euro and the project runs from 2010 till 2015. **PIC**: Here you see some stages in the process, like labeling the folders, **PIC**: separate lines for boxes and folders, **PIC**: the documents entering the scanner, **PIC:** the scan results. This gives you an idea of how the process can work. The RKD is seeking to make use of the experience Naturalis has gained, but the project as a whole will be more modest in scope and its use of resources.

[slide 46]

All of our documentation is organised in these beautiful green boxes. There are about 32,500 of them, and growing. The total length when lined up is about 3,5 kilometer. The estimation of the number of images is between 5 and 7 million. They are organised by medium, period, and country.
PIC. Here we see a box with documentation on paintings about religion from several painters from the Southern-Netherlands (Belgium) in the 17th-century, whose names are added.
PIC. The folders inside a box also have metadata on stickers that has to be processed.

[slide 47]

Handwritten notes with the images can be important art historical sources. It is not our goal to compile all the scanned images into full-descriptive records as artworks like the ones in our Images database. In this project the handwritten texts will not be transcribed. They will be presented 'as is', (the idea for now) but there will be room for tagging and transcribing by our users, so we also want to use crowd sourcing as a way to further index the images.

[slide 48]

The content of the boxes differs greatly, from neatly arranged photographs and cut-outs with typed descriptions, and often also handwritten annotations, **PIC.** To lots of loose clippings and cut-outs, often without any description.

[slide 49]

The digitised documentation collection will be stored in a DAM (digital asset management system) and presented within a new dataset with a limited set of 22 fields that are based on the limited data that is available on the labels and such. We will make use of the data in our existing digital collections to complete information when possible.

[slide 50]

As I mentioned previously, the digitisation was planned to start this year, but the project was postponed because we needed more time to prepare for this far-reaching, fundamental project. Apart from the positive effects it will have on our mission, the effects on our traditional organisation are quite substantial.

[slide 51]

Not all the effects are considered desirable (there is some anxiety about it).

[slide 52]

We are aware that we have to do something with our documentation, or else it will inevitably lose its relevance for the larger public. Our expectation and wish is that this project will start somewhere in the near future. One reason, among others, for the RKD to participate in PHAROS is the development of image search and annotation tools. It would be fantastic to be able to compare digital images and annotations across so many great collections.

[slide 53]

End, contact weda@rkd.nl

[slide 54]

For more information, please contact:

Inge Reist, Project Coordinator, Frick Art Reference Library / reist@frick.org

---

[1] This paper draws heavily from contributions from all of the project collaborators. "PHAROS" is an allusion to knowledge, navigation, and one of the seven wonders of the ancient world. In addition, although we realized a name can be as good as an acronym, the spelling of PHAROS has multiple associations with what we are trying to do, e.g., **PH**(oto)**A**(rchive)**R**(esearch)**O**(nline)**S**(earching) or **PH**(oto archive)**A**(rt)**R**(esearch) **O**(nline)**S**(ite). The reason for not emphasizing the acronym is that ultimately we hope to see the consortium offer a myriad of possibilities for research and mere online searching sells those goals short.

[2] In the last decade a growing group of scholars have been studying the importance of photography as both a tool and a medium of art history. An excellent introduction to this emerging field is the book *Photo Archives and the Photographic Memory of Art History,* edited by Costanza Caraffa and Patricia Rubin. These scholars have also advocated for the Florence Declaration, which emphasizes the importance of preserving original photographs as material evidence.