Digitization of Documentation
at the Scientific Archive of
the National Institute of Archeology with Museum

Katrina Koleva[1], Kalina Sotirova[2], Elena Naidenova[3]

[1] National Institute of Archaeology with Museum – Bulgarian Academy of Sciences
[2] Institute of Mathematics and Informatics – Bulgarian Academy of Sciences
[3] National Institute of Archaeology with Museum – Bulgarian Academy of Sciences

Abstract. This paper is focused on the theory, standards and good practices in digitization of archival documents, as implemented in the Scientific Archive at the National Institute of Archeology with Museum (SA of NIAM-BAS). We explored metadata standards for archival documentation ISAD, EAD and "universal" cultural metadata standard Dublin Core trying to choose the *best fit* for representing of all the fields of SA of NIAM to *de facto* "standard", used in NIAM-BAS since the 60^s. The archival documentation of archeological museum is very specific and contains five different types of documents: scientific report, field diary, field inventory book, photographic documentation and graphic documentation. The results and current annotation specifics are explained. Perspectives for future information system development are outlined.

Keywords: Digitization; Archive; Museum; Information Systems; Catalogue; ISAD; Dublin Core, EAD, CIDOC-ARCH

1   Digitization and Archives

Memory institutions, being repositories of artifacts, information and knowledge, approach digitization in a different way. In Bulgarian archives digitization started around 2006 with the *State Archive Agency*. Currently their repository is the only one, which is almost fully available online and has publicly available *Digitization Guide* [11]. The team of Scientific Archive (SA) at NIAM did a benchmark analysis for comparison of standards *for archival documentation* used in archives, museums and how their element sets correspond to the specifics of SA [6]. The list below shows four archival standards, which were part of our observation:

- **ISAD**(G) – ISAD(G) (General International Standard Archival Description) is a standard from 1994 of International Council on Archives (Canada), and contains 26 metadata in 7 categories. This standard is used by the Archive State Agency and we are using its` six obligatory fields, namely: (1) Reference code, (2) Title, (3) Name of Creator, (4) Dates of Creation, (5) Extent of the Unit of Description and (6) Level of description.

- **ISAAR** (CPF) - ISAAR (CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) is similar to ISAD, but does not fit well the specifics of our documents. Customization is always an issues when choosing metadata standard, but as starting point the common frame should fit document`s specifics.
- **EAD** (Encoded Archival Description) is the standard of the Society of American Archivists and MARC Standards Office of the Library of Congress; its` main focus is description of archives and collections, incl. coding of documents. Resources are described via so called *finding aid,* which differs from traditional catalogue records by its hierarchical structure. EAD was developed as a way of marking up the data contained in *finding aids* so that they can be searched and displayed online. It could be a good choice for standard for cultural institutions which are planning to provide open access to their collections online.

We should mention here **OAIS** – ISO 14721:2003 standard [8, 9] of the Governing Council of Space Data System, since it is crucial for current EU Open Access policy for all public information. Unfortunately it is not applicable fully in the SA at the NIAM-BAS, because of various legal and institutional restrictions. In *University of York* (UK), esp. *Archaeology Data Service* at the *Department of Archaeology*, works hard at the documenting of archaeological information. They use set of metadata schemas depending on the type of information. Most of their data is described using *extended* **Dublin Core**, **DataCite** and **OAIS** (which includes more properties for spatial and temporal information, dissemination and transfer procedure) [9,12]**.** Most of the projects of the colleagues from York University are focused on data sharing, unlikely NIAM, but inter-institutional know-how exchange is going on.

2   Scientific Archive at NIAM

2.1   History, Structure and Policy

The SA has a complex structure, both as administrative unit and as an (not typical) archive. From one side, the Archive is part of the archeology museum, uses thesaurus of archeology science and archival science, while at the same time follows restricted access policy.

SA was created as a separate department at NIAM-BAS in 1973. Its` main goal is to receive, register, keep and preserve the scientific documentation of annually conducted archaeological excavations in Bulgaria. In 2015 in SA was launched a digitization project, whose main aim is creating e-documentation. While the infrastructure, hardware and software were delivered in the new office, the employees took a qualification training course on the digitization of archival documentation. As a result, digitization itself started based on a serious theoretical ground. In addition, due to lack of *Digitization guides for archival documents* (in Bulgarian), the team created several workbooks in order to fill this gap. Among these Guides are: „Concept for Digitization of the Scientific Archive at the NIAM-BAS", "Instructions for Scanning and Description of Archival Documentation", "File Naming Procedure", "Transliteration

Table for the Collection`s Names", which uses the rules in the Bulgarian "Law for Transliteration" [10].

Metadata standards for archives are not the only type of institutional and content standards we took into account while doing benchmark analysis for documenting practices of heritage holders [6, 7]. We added CIDOC-CRM [3] standard for cultural information (ISO 21127:2006), which latest draft version 6.2.1, is encoded in RDFS and follows the requirements of the Semantic Web. Using CIDOC-CRM is recommended by ICOM, although wide spread in Europe is the use of much simpler DUBLIN CORE standard. Another good practice in archeology documentation is the use of CIDOC-ARCH standard (CIDOC International Core Data Standard for Archeological Sites and Monuments), followed by GETTY Institute in its` project "Arches"[1, 2] It defines minimum obligatory information categories, among which are: (1) identification fields, (2) institution, (3) references, (4) object name, (5) description, (6) material, technique, (7) dimensions, (8) shape/form, (9) archeological context, (10) inscriptions, (11) period/dating, (12) conservation status. As for NIAM, CIDOC-ARCH, Dublin Core or ISAD – the choice might be adequate only after further research and field-set customization (look 2.4.).

2.2   Terminology

The structure of the *Digital Archive* follows the structure of the SA, which means that the three main terms in use needed re-defining. They are:
— <**DO**> **Digital Object** refers to *one particular element in a set of 5 elements* of the scientific documentation (scientific report, field diary, field inventory book, photographic documentation, graphic documentation). The different parts (pages) of this element is assembled in one PDF file during digitization.
— <**DAC**> D**igital Archival Collection** consists of all *5 elements* thus representing the scientific documentation of a certain archaeological site in one concrete year; it has unique <digital inventory number>
— <**DIN**> **Digital Inventory Number** is number of 11 digits, which give information about the number of the archival collection at the SA, the year of the archaeological research and a serial number of the digital archive collection.
These three terms are the core of current digitization practice at the Digital Archive.

2.3   Digitization Cycle

Creation of Digital Archive at NIAM-BAS has gone through the following phases, which happened in 2015:
— Selection and assurance of appropriate technical infrastructure, based on examination of archival documents; Preparation of digitization strategy and agenda
— Benchmarking analysis of metadata standards for archival documentation [6, 12];
— Instructions of the team and creating of 6 working documents, which formalize scanning, annotation, storage activities and access policy.
   Digitization cycle we follow can be summarized in the following steps:
1. File naming and folder organization;

2. Setting of the resolution and color of the scanned copies, setting of file formats;
3. Image processing, protection of the PDF files – creation of digital object, adding a page with copyright information (**Fig.1**), watermark, header/footer, the digital inventory number (**Fig.2**);
4. Introduction of metadata (23 fields) and registry of the digital archive collections (5 fields) through filling in a table in MS Excel
5. Quality control of the created digital objects; Ensuring the safety and integrity of the original archive documents.

Fig. 1. Copyright Page   **Fig. 2.** Watermark on the page
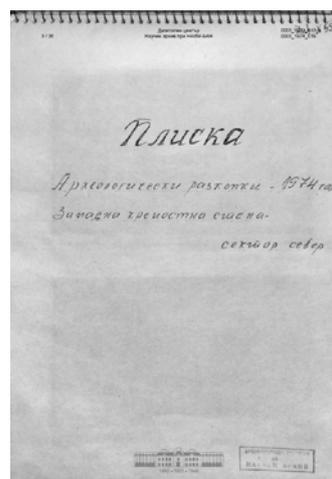
12 738 files of scanned pages have been processed by the team for less than 9 months. This result is very good taking into account that the team consists of two people only, which does not use specialized information system. The scanned documents are described in annotation-tables, in a template. In this MS Excel template (which can easily be converted into database via MS ACCESS) the team uses strictly defined file-naming system, so that searching, filtering and sorting is easy. The future plan includes the use of specialized museum software, based in a good scenario on CIDOC-ARCH standard [1,2]. Using a documentation standard though requires further research on policy, legislation and technical issues.

2.4 Description Fields

Based on SA documenting practice since 60s, the authors created modification of description fields, based on (1) Dublin Core elements, (2) obligatory fields of ISAD (G) and (3) de facto "standard" used in SA. De facto "standard" is a legislation framed procedure of museum documentation, which is used in the last decades not in SA only, but in all archives and cultural heritage holders in Bulgaria. We made benchmarking analysis in which we saw that documentation practices we are using and those in the State Archive at the initial phase of digitization (i.e. in 2010) are similar. They also

used MS Excel for description of scanned documents while waiting for institutional decision for a specific, standard based information system. Resuming the benchmarking we made, 75% of the most spread standard, Dublin Core, set of elements are compatible with our requirements, enabling modification of existing fields with additional information. 11 of them are taken from ISAD (G). Our aim was to find relevance between description of the fields of (1) SA, (2) DC and (3) ISAD (Table 1.).

Table 1. Comparison List of Description Fields

No.	DUBLIN CORE	SA of NIAM-BAS	ISAD (G)
1	Title	**Number of the archival collection**	Title
2	Subject	**Name of the archaeological site**	Title
3	Coverage	**Geographical location**	-
4	Date	**Year of the archaeological excavation**	Dates
5	Creator	**Archeologist (Name of the Head of Excavation team)**	Name of Creator
6	Subject	**Theme, sector**	-
7	Source	**Inventory Number**	-
8	Identifier	**Digital Inventory Number**	-
9	Format	**Number of pages**	Extent and description
10	Type	**Text type of the archival documents** (manuscript or typewriting)	-
11	Language	**Language**	Language
12	**DC: Type** of Archival Document	**Scientific report** (number of pages)	Scope and content
13	**DC: Type** of Archival Document	**Field diary** (dates; number of pages)	Scope and content
14	**DC: Type** of Archival Document	**Field inventory book** (dates; number of pages)	Scope and content
15	**DC: Type** of Archival Document	**Photographic documentation** (number of pages and photos)	Scope and content
16	**DC: Type** of Archival Document	**Graphic documentation** (number of drafts)	Scope and content
17	**DC: Type** of Archival Document	**Notes**	Note

If the goal for NIAM is acquiring uniform information system for all administrative units, incl. museum and SA, Dublin Core can be the basic standard. The table above shows **set of 17 categorical fields** (obligatory and recommended, 23 altogether) for description of scanned documents. The table is to be further developed, but this set of field is final, since it gives full description of the documentation in SA.

3 Future Plans

NIAM-BAS tries to follow internationally recognized rules for protection and preservation of archaeological heritage and documentation, and will continue to work hard for its technology enhanced presentation. The digitization of the Scientific Archive will introduce modern technologies for presentation of museum and library funds of

NIAM. Digitization of museum collections and their presentation in Europeana platform is the next step of the process.

Following EU Directive for open and remote access to all digitized cultural content, we are planning to build an *museum information system for digitization and archeology documentation*. In addition, NAIM plans to coordinate creation of an *e-platform, collecting information from all institutions and structures, involved in conservation, research and management of archaeological heritage*. To mention a few: NIAM-BAS, National Institute for Immovable Cultural Heritage (NIICH), archaeological museums all over the country etc. The platform will include various types of data, so minimum requirements are: (1) description according to CIDOC-CRM or relevant standard, providing interlinking between digital objects, collections and lists; (2) ontology based search functionality, referencing (3) high security level for access; (4) user friendly interface, (5) easy user/system interaction. The aim of such a platform is improving management, protection and promotion of Bulgarian archaeological heritage, as well as facilitating communication between all institutions and structures, involved in research and management of archaeological heritage. Meanwhile, we will continue digitizing the archival documentation, as described here.

References

1. **ARCHES** [http://archesproject.org/standards] Last visited: [12-06-2016] The data relationships that exist in the Arches information system are based on a mapping of CIDOC-CRM to the draft International Core Data Standard for Archaeological and Architectural Heritage (CDS), which is now being finalized by CIDOC, the International Committee for Documentation of the International Council on Museums.
2. **Carlisle,** P. K. at all., The Arches Heritage Inventory and Management System: a standards-based approach to the management of cultural heritage information, 2014 [http://archesproject.org/wp-content/uploads/2014/10/I-1_Carlisle_Dalgity_et-al_paper.pdf] English Heritage, World Monuments Fund, and the J. Paul Getty Trust
3. **CIDOC- CRM** [http://www.cidoc-crm.org/] Last visited: [12-06-2016]
4. **Ivanova, K. at all** (2012) Access to Digital Cultural Heritage: Innovative Applications of Automated Metadata Generation, Published by: Plovdiv University Publishing House
5. **ISO**: A Reference Ontology for the Interchange of Cultural Heritage Information. ISO 21127:2006. [http://www.iso.org/iso/catalogue_detail?csnumber=34424]
6. **OAIS standard** [http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284] Last visited: [12-06-2016]
7. **OCLC Research Centre** [http://www.oclc.org/research/publications/library/2000/lavoie-oais.html] Last visited: [14-06-2016]
8. **State Gazette** [Държавен вестник], issue 68/ 2.08.2013 "Transliteration Table for the Collection`s Names" [http://www.ciela.net/svobodna-zona-darjaven-vestnik/issue/4853/published] Last visited: [12-06-2016]
9. **Technical Guidelines for Digitization of Archival Documentation – State Archive of Bulgaria** [http://www.archives.government.bg/uploaded_files/1_Conception.pdf], []Last visited: [12-06-2016] [http://www.archives.government.bg/uploaded_files/2_Representation.pdf] Last visited: [12-06-2016]
10. **Technical Guidelines for Digitization of Archival Documentation – USA** [http://www.archives.gov/preservation/technical/guidelines.pdf] Last visited: [12-06-2016]