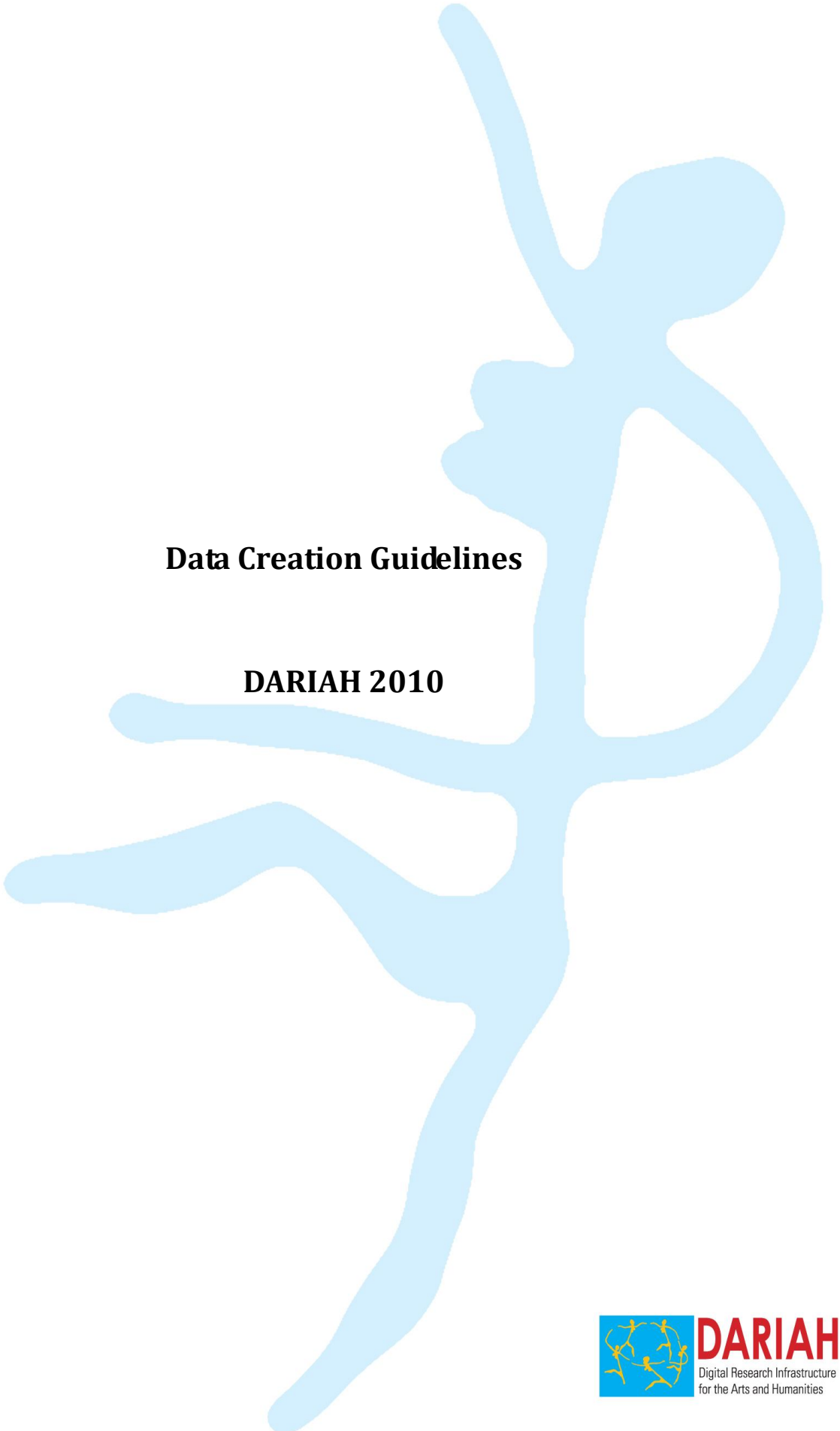
Data Creation Guidelines

DARIAH 2010



DARIAH
Digital Research Infrastructure for the Arts and Humanities

Table of Contents

Executive Summary

As part of the DARIAH strategic guidelines this analysis tackles the main issues of data creation. It provides a set of guidelines to which researchers should adhere when planning a digitization project. It is divided into two parts: the first one refers to data creation principles and the second to the basic steps for the implementation of a data creation project.

This policy is intended to serve as the first stop for someone looking for an overview of principles, as a set of principles for a user ready to provide his content to DARIAH, or a user who visits the DARIAH site for first aid documentation practices. In that sense, we pinpoint the importance of each of the principles explicitly. It is crucial that decision makers and staff fully comprehend the impact of their choices or omissions. Why is intellectual property rights clearance necessary? Why is planning so important? What makes life-cycle

management strategy impossible to leave out? In what way information consistency and quality have an impact on the outcome? Why does interoperability matter?

Additionally, we thought that it would be more practical to provide and describe the basic steps of a data creation project in order to give first hand guidance. We have divided our guidelines into two procedural parts: the first one refers to the decisions that have to be taken during a preparatory phase, such as budget, methodology and IPR issues, user needs, resources assessment, selection of documentation standards, software, hardware and equipment choices, preservation insight and technical considerations. The second part concerns the digital creation processes. Here we deal with preparation arrangements, establishment of workflows, quality control, specifications on the digitisation of all sorts of data, metadata and storage.

Introduction

As more and more scholars and researchers started to implement digital research in their projects, whether in the framework of an institutional policy or as individuals with the increasing use of the Internet and the assistance and experimentation on a growing set of software applications, the digital humanities has grown in scale and importance. In many cases, evolution toward the 'digital' in the humanities occurred in the context of data creation projects. Furthermore, digital scholarship would have been unthinkable if there was no digitized content to analyse and/or share. Therefore there is an ever-growing need to digitize one's content. Whether at an institution, a library, a museum, an archive, a library or an individual researcher there is a growing understanding of the importance of having data available in electronic format. What's more, not only research resources but also research results are being digitized or even, in recent years, born digital.

The first experiments and successful projects have already taken place and the lesson learnt is that we need to systematise the digitisation process in an effort to communicate our sources effectively and sustain them for the long term. Consequently, one of the main priorities in a digitisation project is to ensure the consistency and sustainability of the new digital data formats and the successful management of the data creation process itself. In this data creation policy we will sketch out the guidelines as to the planning, implementation and management of data to be digitized. The guidelines concern the consultation of a set of standards and best practices already tested and promising adequate sustainability.

Before proceeding to the data creation stages we need to define what digital creation means under the perspective of this analysis. Digital creation is the transformation of a physical or analogue type object into a new digital representation form. By "digitisation" we mean any process by which information is captured in digital form, whether as an image, as textual data, as a sound file, or any other format. As such, the creation of data requires the employment of skilled personnel and the utilisation of a certain methodology, process planning and the use of appropriate equipment and software tools. This short policy presents the criteria involved in this fundamental process and is not an extensive description of every single technical parameter required in every single case.

Audience

The DARIAH Collective Intelligence is envisaged as a primary resource for all European scholars working in the digital arts and humanities. Among the features that the information aspect will support and provide are guidelines to good practice. As part of this information layer this report provides information on data creation for arts and humanities researchers.

Consequently, the audience for this document belongs to two levels. This document is addressed to the DARIAH partners and communicates strategy regarding policies, while, at the same time, it can be used as an introductory report on digital creation guidelines within the information facet of the DARIAH infrastructure. The information facet is addressed to the researchers' environment (research actors, research infrastructure actors, supporting actors) as well as to digitisers, digital humanists and Cultural Heritage professionals.

Aims

In recent years, best practices and guidelines have gained the appreciation of organisations and communities and many successful efforts have been made for publishing such guides. In most cases these policies concern needs of specific disciplines or institutions. Rather than produce another guideline, within the scope of the DARIAH preparatory phase, this report aims to summarise and highlight the most significant parts of the digitisation process. It is a presentation of data creation processes that will be useful to resource creators, users of the DARIAH infrastructure and arts and humanities scholars in general.

Scope

Seen from the perspective of data creation, this set of guidelines lays emphasis on relevant methodologies and briefly touches upon preservation, management and IPR issues. While there are overlaps with the other policy documents in the DARIAH strategic framework (Collection Scope and Coverage, Collection Ingest, Management and Preservation,

Compliance as a Trusted Digital Repository), these guidelines are focused on the requirements for the successful completion of a data creation project.

Basic Principles

Experience in digitisation projects has produced a series of principles. Failure to observe them in a project will most certainly result in poor results. The key principles are as follows:

- Intellectual Property Rights clearance is the first thing an organisation has to consider before proceeding any further, as failure to secure rights clearance may put an ambitious digitisation project at risk.
- Planning the procedures and the usage of the resources is an additional prerequisite. Its value is self-evident and puts things into time perspective while at the same time helps to assess an organisation's resources and the management of the process.
- Life-cycle management refers to all stages of creation and maintenance of the data.
- Communication and sharing of digitized data is closely related to the adequacy of technical specifications that allow interoperability.
- Sustainability is a crucial principle because, if applied, it safeguards the future of the project and its results.

During selection processes and decisions concerning equipment, software applications, standards and skilled personnel, it is critical to select the best available solutions. If all these requirements are met, even a small budget project may have long-term use, development and great impact both for the organisation and for its audience.Some of the basic principles described may coincide with the steps taken in the life-cycle management of data creation.

Planning & Implementation

Project planning

Even though the stages of a digitisation project do not follow a linear progression, we will try to unfold the process in a logical manner. The planning of a digitisation project takes into account the aims of the project, its coverage, its collaborations, and its sources. Some of the basic questions that have to be answered during the preparation of the project planning are:

how will the standards, methodologies and specifications of the digitisation process be determined; who will be doing what (human resources); what will be the work location/s and what is the estimated duration for each stage and for the whole endeavour.

It is crucial for the viability of the project to ensure that good communication is established among partner organisations and/or individuals. Changes of plans occur often, especially when it comes to requirements and participants should be able adjust to iterations along the way. Good communication is necessary, while meetings between co-workers at close intervals for the assessment and confirmation of the work-flow practices and short-term results are strongly recommended.

Evaluation of user needs and audience access modes

Some of the key decisions that have to be made are on the types of users the project is addressed to and the services the organisation has decided to provide (access, rights to share and reproduce). The project's aims are closely inter-woven with its audience and the subsequent impact on scalability and access modes. If the organisation decides to allow the access and use of its surrogates by its audience, there has to be a detailed plan on the rights management, the information architecture, the implementation, the security and the sustainability of such choices. A well-developed authentication system and a corresponding user interface design are necessary to support effective search and retrieval of content.

Selecting data sources

The outcome of a digitisation project should have an added value for the organisation and conform with its overall strategy. This kind of assessment will enable those responsible for the planning stage of the project to state the aims of the digitisation initiative and define the selection criteria of the material. Some general aims that may affect the selection criteria are: replacing analogue with digital formats or replacing obsolete digital formats with new ones ensuring preservation, enhancing access to information, providing new or additional documentation to objects, meeting the interests of users including researchers, students, general public or any other type of stakeholders, enabling access to metadata and interoperability with other projects.

One of the most important criteria to guide the selection process is that of users' requirements. It is essential to know the user group of each organisation and potential digitization project, and the availability of equipment and skilled staff to attain the target. It would be also wise to predict what users might need in the future or which new users the organisation would like to attract.

While selecting the sources one has to decide whether the digitisation will be based on the original object or on an existing surrogate. Such criteria as access and the peculiarities in handling play a significant role in the final selection decisions of both the material and the equipment (for example, special equipment might be needed for an oversized original artwork or a preserved fragile book). Using an intermediary can frequently be the best solution since scanning the originals can increase the cost. Once choices are made and stated, it will be convenient for the subsequent workflow stage of the project to generate a catalogue of the items to be digitized with a description of their main characteristics. This step will help to determine to which work process each item belongs.

Clearing rights

Prior to any digitisation project organisations should establish the copyright status of their source material. This, of course, requires research to establish how many of the source material rights are held by the organisation and for which objects further investigation is required. Accordingly, the working group will decide which items of the source material will not be digitized and for which and with what criteria it is worthwhile to secure permission for digitisation. The same of course applies for the derivatives of the original work to be digitized. Once permission for digitisation is obtained, then comes the procedure of requiring permissions to use. The usage rights may concern not only the right to digitize a resource, but also what one can or cannot do with the digital version, if and how one can disseminate it and make it accessible to the public. After rights clearance is obtained, the conditions of use should be included in the metadata of the digital format and made visible on display on the Internet or any other form of publication.

The entire IPR clearance procedure is a project by itself. It is therefore advisable to establish project planning as well as risk management to the material to be selected for digitisation. The application of copyright rules may vary from country to country. Apart from individual items, databases as software systems, which organize a collection of items, are also under a scheme of rights and their status and the manner in which they can be used for each project should also be clarified. The most formal way to address the issue of copyrights management is to apply a contract to be called “License agreement” comprised of terms and permissions between all associated parties (creators, rights holders). This set of rules are set taking into account the legal framework, the creation history of the material, the individual rights people hold and the individual license agreements of the right holders or others may issue.

Selection of standards

Use of standards is the most effective way to achieve interoperability, consistency and sustainability for an organisation's content. There are different kinds of standards in the sense that they apply to different purposes and levels of information. Typically there are three levels of information: the description of digitized material (structural metadata: METS, SMIL), its administrative information (IPR rights, location) and the description of the object depicted (CDWA, AAT, Dublin Core, CIDOC). As for the structural metadata of the digitized objects the following are the most popular types of data and corresponding standards: Text (XML, EAD, TEI), Audio (WAVE, AIFF, MPG3, RA), Video, (MPEG1-3), 3D Video (MPEG-4) and Image (BMP, PNG, TIFF).

Each of these standards has its own advantages and weaknesses for documentation and appeal to user groups and addresses specific needs according to the project's aims. On all sets of information the organisation may want to apply restrictions to the use and display to individuals or certain groups of users. Criteria like the type of object, the depth of documentation and the willingness to share it will guide the selection through the existing documentation standards.

Selection of software and hardware infrastructure

The infrastructure needed to carry out a project typically consists of equipment, operating systems and software for image creation and manipulation, data authoring, data managing software that index the content and servers to support network clients. Additional consideration should be given for storage devices such as local hard drives and storage servers.

The medium, format, size, and fragility of the original material are among the primary factors affecting equipment choice. For text documents scanners suitable for the material should be selected. Digital cameras, with appropriate complementary tools are a versatile option for bound material. Audio and moving image materials present their own problems for digital capture due to the variety of source formats (records, cassette tapes, PAL and NTSC formats) and the difficulties in accessing analogue devices. The digitisation of objects for 3D representation is achieved by applying special methods for moving images formats (QuickTime VR) and requires the modeling of the object with respect to features and dimensions (CAD/CAM applications).

Digital creation

Digital creation is a composite process and requires the assessment of the source material and the establishment of procedures for the actual digitisation. It is essential to organize and manage the material according to its format (paper, paintings, slides, negatives, transparencies, maps, postcards, etc.) and its properties (fragility, size, access etc.). This knowledge will determine the selection of methods and equipment as well as the selection of specialized staff. Having accomplished this step it is easier to establish the most effective workflow.

It is also necessary to establish the minimum capture requirements for the entire collection, depending on the original material, the resources and the intended use. Sample testing is advisable to provide a benchmark for sampling rate and precision.

Quality control

One of the critical steps in a digitisation workflow is that of quality control. When checking digital images for quality one usually checks the correctness of image size, resolution, file format, contrast, and orientation. The work environment is of great importance too, since it affects the quality of the captured images. Equipment adjustments, appropriate software, steady and small group working in the process, and making use of checklists for the procedures may safeguard the overall quality of the project.

Storage and preservation

The issue of storage interrelates closely with the issue of software and operating systems used for the digitisation process (delivery, back-up, etc.). As all these are under constant development and change, storing media and file formats becomes a tricky process. The storage requirements have to be set in every digitisation project to secure maintenance of access to the digitized content. Furthermore, storage devices should support storage demands regarding types of files, formats and size.

Most common storage options are hard, optical (CD-R, DVD-R), tape (DLT, AIT) and networked drives (RAID), which also need to be backed up onto other storage formats at close intervals. It may also be worth considering storage on servers and repositories.

Since the usability of data depends on operating systems and functional software, it is advisable to disengage the content from the discovery and display systems. The formulation of a preservation strategy plays a decisive role in ensuring the sustainability of any project.

Migration to new formats at regular intervals is usually inevitable and is adopted as the best solution to follow when working environments change. A successful policy involves frequent checks on the data to ensure that they are still readable and on the media to ensure that they have not become obsolete and that they are replaced on time. Perhaps the only way to ensure accessibility to content is the use of standard data formats (SGML, XML) that keep the data intact and portable. In general, following standards can keep an organisation's resources interoperable and sustainable over time. Maintaining metadata of the digital files is equally crucial to preservation of digital resources, while their sustainability depends on the medium and the standards used (ASCII, Unicode, XML).

However, preservation also concerns the physical deterioration of digital media. The physical location that is used for storage should meet humidity and temperature requirements. Likewise, the physical media on which the digital material is stored need constant checking, appropriate handling and a suitable environment. The best risk management practice is to store a second full set in another location in case of theft, fire or other natural disasters.

Links to guides, standards and practices

CLIR: Council on Library and Information Resources: "The projects and activities of CLIR are aimed at ensuring that information resources needed by scholars, students, and the general public are available for future generations." http://www.clir.org/

DLIB Forum: "The D-Lib Forum supports the community of researchers and developers working to create and apply the technologies leading to the global digital library." http://www.dlib.org/

LOC: Library of Congress: "The Library's mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." http://www.loc.gov/

NINCH: National Initiative for a Network Cultural Heritage: "A coalition of arts, humanities and social science organizations created to assure leadership from the cultural community in the evolution of the digital environment." http://www.ninch.org/

RLG: Research Libraries Group: "The Research Libraries Group, Inc., is a not-for-profit membership corporation of universities, archives, historical societies, museums, and

other institutions devoted to improving access to information that supports research and learning." http://www.rlg.org/rlg.html

PADI: "The National Library of Australia's Preserving Access to Digital Information initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access." http://www.nla.gov.au/padi/

HEDS: Higher Education Digitization Service: "The Service provides advice, consultancy and a complete production service for digitization and digital library development." http://heds.herts.ac.uk/

TASI: Technical Advisory Service for Images: "Advise and support the academic community on the digital creation, storage and delivery of image-related information." http://www.tasi.ac.uk/

AHDS: Arts and Humanities Data Service: "Create and preserve digital collections in all areas of the arts and humanities." These pages are no longer updated but still contain some useful information: http://ahds.ac.uk/