Please cite this as: Abele, J., Dürr, A-M., Berger, S. and Schrickel, M. 2026 'Digital Archaeological Primary Documentation Data in Baden-Württemberg (Germany). The LAD-BW's Path from Standards to Archives', Internet Archaeology 72. https://doi.org/10.11141/ia.72.5

Digital Archaeological Primary Documentation Data in Baden-Württemberg (Germany). The LAD-BW's Path from Standards to Archives

Jonas Abele, Anna-Marie Dürr, Steffen Berger and Marco Schrickel

Summary

The paper presents the current state of long-term data storage for archaeological primary documentation data at the Landesamt für Denkmalpflege Baden-Württemberg (LAD-BW). A central long-term repository was established in 2016 due to the significant increase in digital primary documentation data, driven by a rise in construction activities and associated rescue excavations in recent years, with the latter increasingly being carried out by archaeological service companies since 2016. Since 2018, guidelines have defined data structures and formats, and a central project identifier. Metadata and essential information (e.g. excavation reports and selected photographs) provided in the ADABweb information system ensure internal findability. The repository currently contains more than 6400 data packages (70 TB, as of July 2025).

Data analyses revealed heterogeneous data structures (particularly from 2000–2017), ongoing format diversity, and storage usage of which roughly 90% is accounted for by image data (JPG/TIFF). One major focus concerns the handling of structure-from-motion (SfM) data, with large volumes of JPG raw data and Metashape project files being deposited between 2018 and 2021; the latter are proprietary and not preservation-ready. Within the framework of a 'legacy data project', selection, data hygiene and migration are being implemented under clearly defined rules and documented as standardised processes.

Outlook: Revising the excavation guidelines and developing new target systems (including a geodatabase) aims to facilitate the creation of guideline-compliant data and strengthen the FAIR (Findable, Accessible, Interoperable and Reusable) principles, particularly through standardised, centralised metadata and an interconnected knowledge base. Two open access publication series (project 'Pilotprojekt In-wertsetzung Ausgrabung (PIA)') accelerate the dissemination of excavation results and, in part, the publication of associated research data.

1. Origin and current status of long-term data storage

An important task of the State Office for Cultural Heritage Baden-Württemberg, Germany (Landesamt für Denkmalpflege Baden-Württemberg, LAD-BW), is to record, document and research cultural monuments in the state. This role includes the implementation and supervision of various archaeological fieldwork measures, such as surveys, trial trenching, excavations and geophysical surveys. The largest share of such activities consists of large-scale rescue excavations and trial trenching in advance of construction projects, as well as construction supervision, usually conducted under considerable time pressure. Within the

framework of these activities, primary documentation data are generated, which today exist almost exclusively in digital form. This continuous stream of digital data poses several challenges for the LAD-BW regarding data management, including implementing standards and guidelines to ensure archivable and reusable datasets. The following sections provide a practice-orientated account of the current state of the data management strategy, including a project within the framework of NFDI4Objects — the National Research Data Infrastructure for the Material Remains of Human History — that addresses the handling of non-standardised digital legacy data, combined with an outlook on future developments.

Since 2005, an internal data backup service has been offered for digital primary documentation data at the LAD-BW, with the prospect of transferring these to a genuine long-term archive (Bibby 2021). Following structural reform in 2015, the headquarters of the LAD-BW (near Stuttgart) gained a more central role, while branch offices continue to exist in Tübingen, Freiburg, Karlsruhe, Hemmenhofen, and Konstanz. This reform created the structural basis for transferring the previously highly decentralised storage of primary documentation data — which, despite the internal backup service, was generally conducted autonomously by the respective data producers — to a central data repository operated by the state IT service provider BitBW. The long-term storage of primary documentation data of archaeological heritage management was initiated in 2016. Since then, not only current primary documentation data but also all available digital documentation from previous years have been successively incorporated. In addition to the unification and centralisation of the storage strategy through the establishment of the central long-term archive, there is another essential reason for a data management strategy: beginning in 2016, an increasing number of surveys and excavations at archaeological sites has been conducted by private archaeological companies, expanding the circle of data producers. Since 2018, guidelines (Firmenarchäologie n.d.) have governed the structure of these data, from folder hierarchies to file formats, with explicit consideration of long-term archiving. A unique identifier (year + sequential number, which must be used) is centrally assigned and embedded in the data — both in the folder and file names and in the attributes, for example in the metadata or in context records. This practice ensures datasets can be assigned to a specific project and uniquely identified. Of particular importance was the introduction at this time of the ADAB folder (Bibby 2021). This folder contains the core information for each project (spatial boundaries as a geodata file, metadata, a selection of images, and a PDF report) and makes them available in the LAD-BW's information system (ADABweb, n.d.). This system ensures the data are accessible and searchable internally. Moreover, the introduction of the guidelines laid the foundation for a structured data management approach that has been further developed in subsequent years and is outlined below.

2. From project to data package – an overview of data management

The focus of the following account is the standardised data package and the associated processes of data management. The package contains all the primary documentation data of a project deemed worthy of archiving. The organisational workflow of data management is illustrated schematically in Figure 1. The majority of the data are generated during on-site documentation and in subsequent postprocessing.

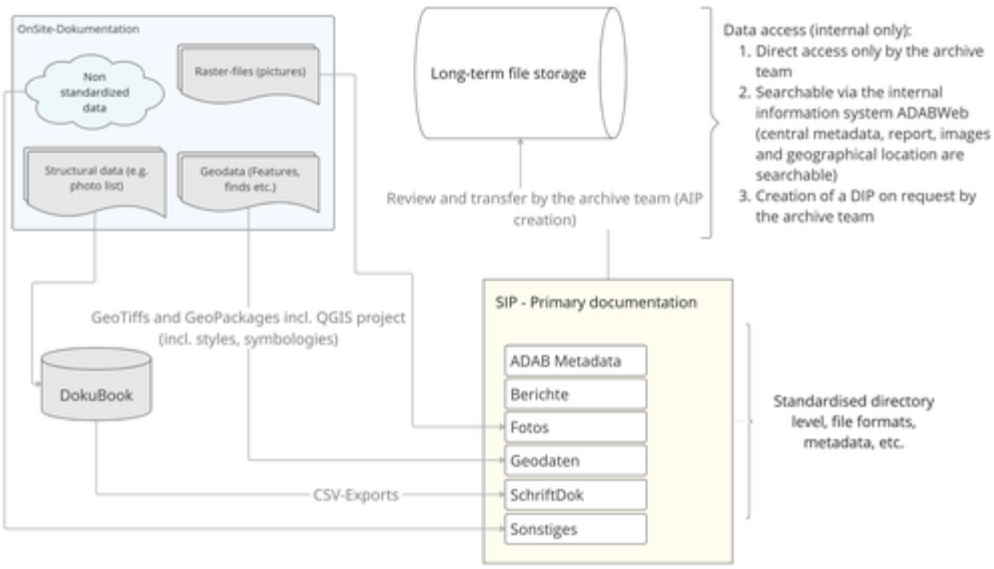

Figure 1: Data-management plan for the primary documentation data of archaeological projects at the Landesamt für Denkmalpflege Baden-Württemberg (LAD-BW). Graphic: Jonas Abele.

2.1. Description of the data package according to the guidelines

At least the data resulting from excavations, trial trenching and construction supervision are largely standardised by the excavation guidelines (Firmenarchäologie n.d.). These regulations define, for example, the structure of the geodata — such as context geometries, find points and georeferenced orthophotos — and their uniform integration into a QGIS project. Within the Submission Information Package (SIP), these geodata are stored in a folder named 'Geodaten'. Structured textual data, such as photo lists and context (feature) registers, can be generated using the excavation database known as DokuBook (see Figure 1) or imported into it via CSV tables. This process ensures the data conform to the guidelines. Within the SIP, these structured data are consistently stored as CSV files in a folder named 'SchriftDok'. All non-georeferenced photos are stored in the 'Fotos' folder and documented in a photo list containing relevant metadata (e.g. photographer, date of capture, context numbers). A report on the project, in PDF/A-2 format, is deposited in the 'Berichte' folder.

In addition to the standardised data covered by the guidelines, every project inevitably produces a set of information and data considered worthy of archiving but not addressed by the excavation guidelines — either because they are too project-specific or simply because they were not defined in advance but nonetheless were created to add value. These data may also be deposited in the SIP by the data producers. For this purpose, the folder 'Sonstiges' ('Miscellaneous') is provided. Practice has revealed, however, that precisely these data, perceived by the data producers as worthy of archiving but not covered by the guidelines, can create certain difficulties within the storage structure. For instance, hand sketches may be scanned or geodata generated that go beyond the guidelines (e.g. interpretative entities within the geographical information system (GIS)). These data are without doubt archivable but, at least in the case of geodata, are usually linked in an interoperable way and thus rarely stored in the 'Miscellaneous' folder but in the relevant geodata folder. It is partly for this reason that the revised Version 4 of the guidelines, which is currently being implemented, foresees an overhaul of both the folder structure and the

handling of archivable but non-standardised data. In particular, the handling of geodata will, in future, be regulated by more comprehensive GIS guidelines.

For the current data management plan, a package of primary documentation data is transferred at the end of each project to long-term storage as a format-based deposit in the form of a SIP.

2.2. Ingest

The ingest process involves several steps. In cases in which a project is conducted by an external service provider, the transfer of finds and documentation to the LAD-BW is accompanied by a formal review of the data package. The main focus of this review is on compliance with the guidelines at the level of data structure. Content review is conducted on a spot-check basis but, given the volume of data and the number of data packages submitted each year, it cannot be carried out comprehensively.

The pre-checked SIP is then transferred to the archive by the data producers and incorporated into the long-term repository. The primary documentation data from projects conducted directly by the LAD are generally passed to the archive team without an intermediate check. Within the ingest process, the ADAB folder (Bibby 2021) plays a decisive role. This folder contains the spatial boundaries of the project as a geometry file, the metadata (Firmenarchäologie n.d.), the project report, and a selection of representative reference images. As a rule, this folder is compiled by the data producers. If the folder is not provided, it is created during the ingest process. The ADAB folder thus constitutes a kind of summary of the essential project information and is incorporated into the ADABweb information system, making the project data accessible and searchable internally. The archive team is responsible for transferring this information to ADABweb. The Archival Information Package (AIP) is deposited in a central repository. At this stage, the archive team generates additional metadata, such as information on the data producer and technical metadata (e.g. a complete file list, including file extensions and SHA-256 checksums). Access to the AIP repository is restricted to members of the archive team, and data are released only upon request.

Orientation towards the Open Archival Information System (OAIS) reference model is primarily at the conceptual level. The long-term repository largely consists of retrospectively aggregated datasets originating from more than 35 years of (partly) digital excavation practice, the creation of which largely occurred without centrally binding standards or guidelines. Therefore, the data can only be understood to a limited extent as OAIS-compliant AIPs. As the majority of the data were generated without the application of binding standards or systematic consideration of archival criteria, their long-term viability, and in some cases, even their archival value, cannot always be assumed (see the data analysis sections, Section 3).

3. Review and analysis of data storage

The long-term repository contains more than 6400 data packages, with a total volume of 70 TB (as of July 2025; see Figure 2). A significant increase in the number of projects is clearly evident following the establishment of commercial archaeology from 2016 onwards. In addition, the strong wave of construction activity between 2020 and 2022 led to a substantial increase of data. Since 2023, data growth has declined somewhat. However, there is often a time lag between the execution of fieldwork and the submission of documentation to the central repository; particularly for larger projects, an extended postprocessing phase is required before the data can be archived.

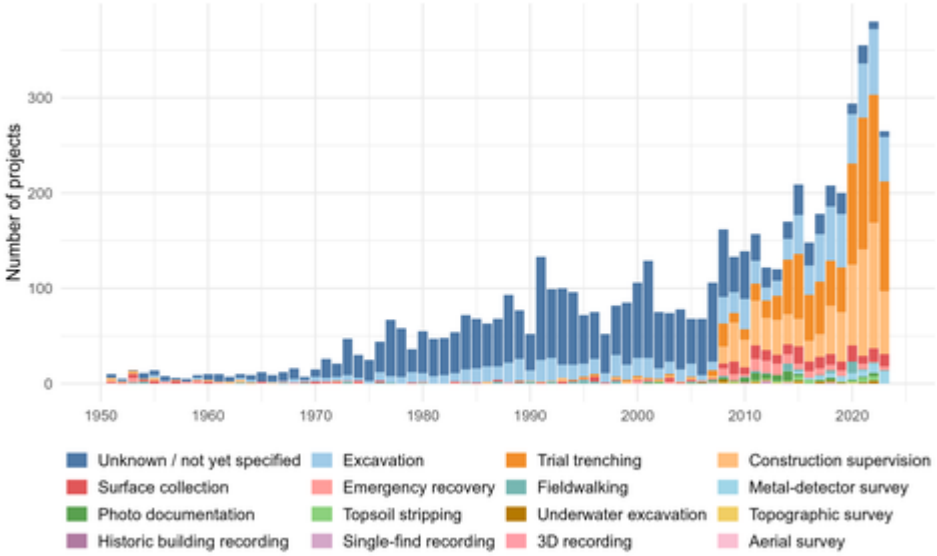

Figure 2: Number of primary documentation data packages stored in the long-term data storage, broken down by year and type of fieldwork activity. Graphic: Jonas Abele.

All data packages are described with the same set of metadata defined in the guidelines introduced in 2018, which include the type of fieldwork activity. For this purpose, a thesaurus is provided (Firmenarchäologie n.d., download section 'Thesauri'). For projects conducted before 2018, such metadata are unavailable and are being retrospectively added. Given the large number of data packages, this task is ongoing and must be accomplished alongside daily operations. Based on the available information, the majority of documentation data stem from archaeological fieldwork, mainly trial trenching, construction supervision and excavations (Figure 2).

The older digital data from between 2000 and 2018 (when projects were increasingly documented exclusively in digital form but without central guidelines) are very heterogeneous regarding their data structures. Of considerable importance, therefore, are the common metadata, which represent the lowest common denominator in information capture and make it possible to render the datasets searchable and retrievable at the metadata level.

To a lesser extent, the long-term repository also contains projects originally documented exclusively in analogue form, which have been partially or fully digitised (Figure 2). Insofar as these projects produced primary documentation data, these data are assigned a project number and can be deposited in the project data archive. The retrospective digitisation has been predominantly project-based and not implemented uniformly. The dataset has grown historically and can no longer be assessed solely at the level of individual projects but must be analysed primarily using quantitative methods. Since 2023, regular statistical assessments have been conducted for this purpose, and some results are presented below.

3.1 Data analysis: archival suitability of the data – historical overview

As part of the data survey, a complete list of all the file types present was compiled and prioritised according to frequency (as of August 2025, the long-term repository contains 10,146,278 individual files). The archival suitability of the individual formats was assessed based on established references (Archaeology Data Service n.d.; IANUS LZA-

Empfehlungen n.d.; Open Preservation Foundation 2022; Swedish National Data Service n.d.). File types with fewer than 400 occurrences were not examined in greater detail and were provisionally classified as unknown. In total, 370 formats could be evaluated.

The resulting statistics, however, are only of limited validity and should be understood as a preliminary assessment, as for core formats, such as Tagged Image File Format (TIFF) or Portable Document Format (PDF), the specific format variants (e.g. PDF/A, TIFF 6.0) were not identified using tools such as DROID.

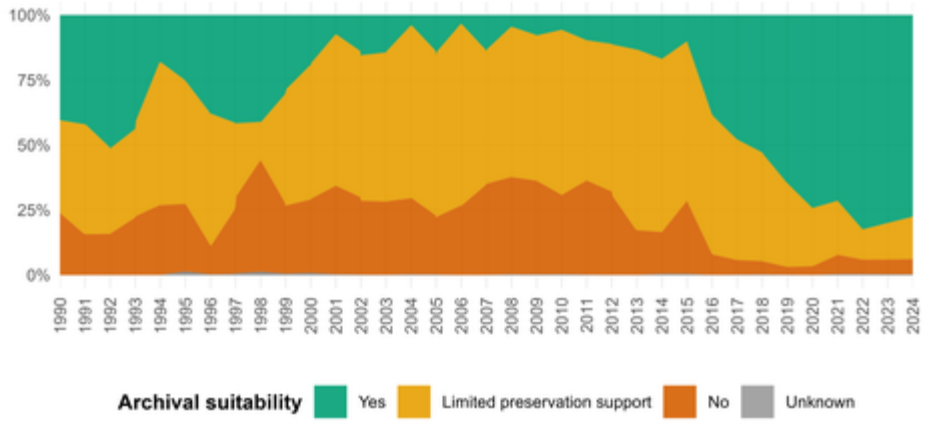

Figure 3: Archival suitability of file formats in the long-term data storage of primary documentation data for archaeological projects at the Landesamt für Denkmalpflege Baden-Württemberg (LAD-BW). Graphic: Jonas Abele.

Based on the references for the archival suitability of file formats, the complete list of data was annotated accordingly, and the annual percentage distribution was calculated (Figure 3). The impact of the excavation guidelines introduced in 2018 is evident, as these included, for the first time, central and binding requirements for file formats: always with a view to their suitability for preservation. Nevertheless, a proportion of only conditionally archivable and even non-archivable formats remains. Among the conditionally preservation-ready formats are, for example, Joint Photographic Experts Group (JPG) image files generated for structure-from-motion (SfM) workflows, which are still occasionally submitted. A particular challenge is posed by QGIS files, including not only project files but also style and configuration files. At present, these files are stored within the SIPs, in part to facilitate the reuse of complex, interoperably linked, and visually prepared geodata. However, a sustainable archiving strategy for these files has yet to be developed and will be an important challenge in the coming years (VLAK-AIS 2024, 9). Especially critical is the period between 2000 and 2017, during which the share of file formats suitable for preservation declined dramatically, resulting in many non-archivable and only conditionally archivable files.

3.2 Data analysis: diversity of file formats

In the second step of the data analysis, a quantitative approach was again applied at the level of file format types to examine problematic formats and their contexts of origin in greater detail. One useful indicator of heterogeneity — and of the resulting challenges for preservation planning — was the number of different file types within a single project. In addition, the number of different file types across all projects of a given year was calculated.

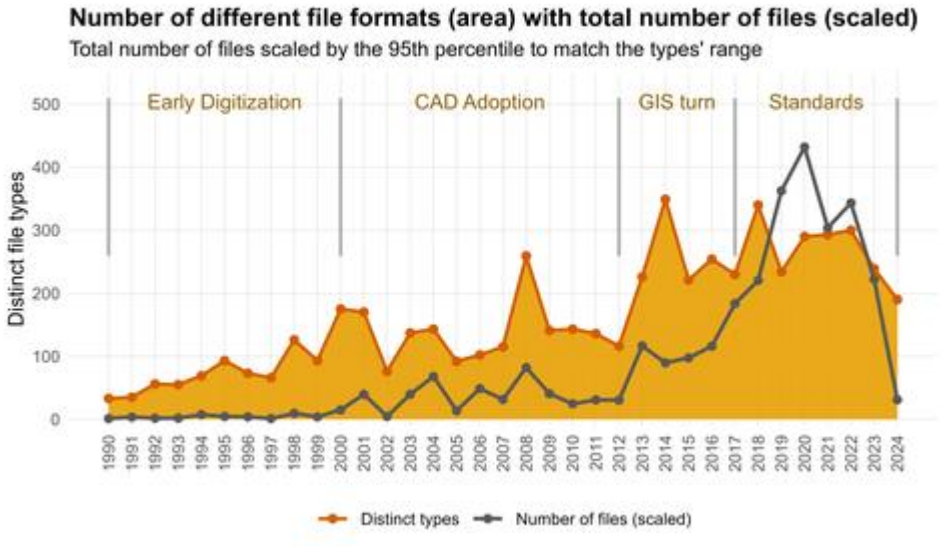

Figure 4: Number of file formats within all projects in a year and total number of files within a year (black line, scaled illustration). Graphic: Jonas Abele.

These numbers can be illustrated graphically (Figure 4) and prove particularly insightful for understanding the genesis of the data collection. In the past 35 years, four distinct phases can be identified.

Phase 1: analogue-to-digital beginnings (1990–1999)

The most frequent file formats listed below reflect the first steps in the emerging digitisation of essential workflows of primary documentation, especially photographic and written documentation. A larger number of Worksheet (WKS) and Writer, Presentation, Spreadsheets (WPS) files indicate the increasing digital recording of lists and reports using Lotus 1-2-3 and Microsoft Works. Digital photography was already in use, though the majority of drawn documentation remained analogue. Overall, this was a phase of digital pioneering. From an archival perspective, the large number of proprietary, non-archivable file formats in text and spreadsheet processing highlights a pressing need for action. Taking action is also necessary because some programmes, such as Lotus 1-2-3, are no longer officially supported by the manufacturer (Wikipedia 2025) and have long since fallen out of use.

List of the most frequent file formats in Phase 1:

- .jpg, quantity: 45,042
- .tif, quantity: 23,888
- .pdf, quantity: 9273
- .wks, quantity: 7981
- .cr2, quantity: 4403
- .wps, quantity: 2771
- .dat, quantity: 2013
- .doc, quantity: 1931

Phase 2: CAD adoption (2000–2012)

With the widespread introduction of computer-aided design (CAD), analogue hand-drawing was systematically replaced in this phase by digital plan production. Drawing (DWG) files represent the second most frequent file type during this period (see the list below). The formats PPB and PRK are internal project files of PhoToPlan, which attest to a key methodological innovation: photographs were rectified on a large scale and integrated into the CAD system either as colour information or as a basis for redrawing. This phase marks the final transition to fully digital documentation. Workflow consistency increased, and the number of digital data packages and files also grew steadily. At the same time, the first standardisation efforts were implemented, such as the development of internal CAD guidelines and the introduction of a central folder structure (Bibby 2021).

From a long-term preservation perspective, all file formats associated with CAD documentation must be regarded as fundamentally problematic. Although these files are mostly not preservation-ready, they are nonetheless of the highest archival value. From the viewpoint of preservation planning, this area requires urgent action.

List of the most frequent file formats in Phase 2:

- .jpg, quantity: 837,153
- .dwg, quantity: 71,445
- .tif, quantity: 68,884
- .dat, quantity: 65,989
- .ppb, quantity: 51,772
- .bak, quantity: 45,857
- .prk, quantity: 45,142
- .doc, quantity: 37,129
- .nef, quantity: 35,892
- .bmp, quantity: 33,943

Phase 3: Innovative transition – GIS turn and SfM (2013–2017)

The process of digitisation became increasingly innovative and diverse during this phase. New tools were tested, adopted, and in some cases abandoned again (Bibby 2021). Accordingly, the range of file formats in use was considerable (Figure 4). Central to this period was the transition from CAD to GIS, which also left its mark at the file level. The peak of this transition occurred around 2014, when Shapefile (SHP) and DWG files were nearly equal in number. Some projects were conducted in parallel using both CAD and GIS, and gvSIG, ESRI ArcGIS and QGIS were trialled. Ultimately, QGIS prevailed, with SHP and GeoTIFF becoming the principal exchange and preservation-ready formats.

List of the most frequent file formats in Phase 3:

- .jpg, quantity: 1,001,506
- SHP (.shx, .shp, .dbf, .prj), quantity: 250,808
- .tif, quantity: 83,881

- .pdf, quantity: 50,882
- .dat, quantity: 46,721
- .qml, quantity: 39,370
- .nef, quantity: 36,832
- .qpj, quantity: 34,208
- .txt, quantity: 27,510
- .log, quantity: 19,370

Phase 4: Consolidation and standardisation (since 2018)

With the introduction of the excavation guidelines, standardisation became more firmly established during this phase; from an archival perspective, this phase also led to a stabilisation of the file formats in use. Despite the sharp increase in the number of projects and files (see Figure 2 and Figure 4), the diversity of formats remained largely constant. The very high number of SHP files (consisting of SHX, SHP and DBF) can be explained by the fact that, up to Version 3 of the excavation guidelines, a strongly file-based representation of archaeological entities was prescribed, for example by creating a separate feature SHP for each stratigraphic unit. The guidelines are currently being revised, with the previous approach being replaced by an object- or attribute-centred recording of this information at the data level. With the planned shift to GeoPackage as the preferred file format, this revision is expected to significantly reduce the absolute number of geospatial files, improving the possibilities for quality control during ingestion.

List of the most frequent file formats in Phase 4:

- SHP (.shx, .shp, .dbf, .cpg, .prj), quantity: 2,640,549
- .jpg, quantity: 1,269,807
- .qml, quantity: 728,903
- .tif, quantity: 341,548
- .cpg, quantity: 316,856
- .qpj, quantity: 261,591
- .log, quantity: 193,618
- .pdf, quantity: 174,255
- .txt, quantity: 155,356
- .gva, quantity: 65,450

3.3 Data analysis: storage consumption

In recent years, not only have the number of projects (Figure 2) and the number of files (Figure 4) increased significantly but also the average project size has grown continuously. This change is hardly surprising, as individual files, such as photographs, have steadily become larger over time as hardware has improved. This point also applies to orthophotos generated using SfM, which produces larger datasets than alternative methods. The annual average project size remained almost entirely below 5 GB until 2012, and a peak of nearly

45 GB was reached in 2019: a ninefold increase compared with the period from 1990 to 2012.

As project sizes increase at the storage level, data packages become more difficult to manage: the creation of Dissemination Information Packages (DIPs) involves long copying processes. Storage costs have also risen considerably in recent years due to the growth in data volume, creating a clear need for action. The objective of this analytical step was therefore to identify particularly storage-intensive file formats and information domains to establish the priorities for potential data-cleansing measures avoiding holdings that are not worthy of preservation.

Unsurprisingly, almost 90% of the occupied storage space is accounted for by image files, mainly JPG and TIFF formats (see also the file lists of the individual phases of data creation above).

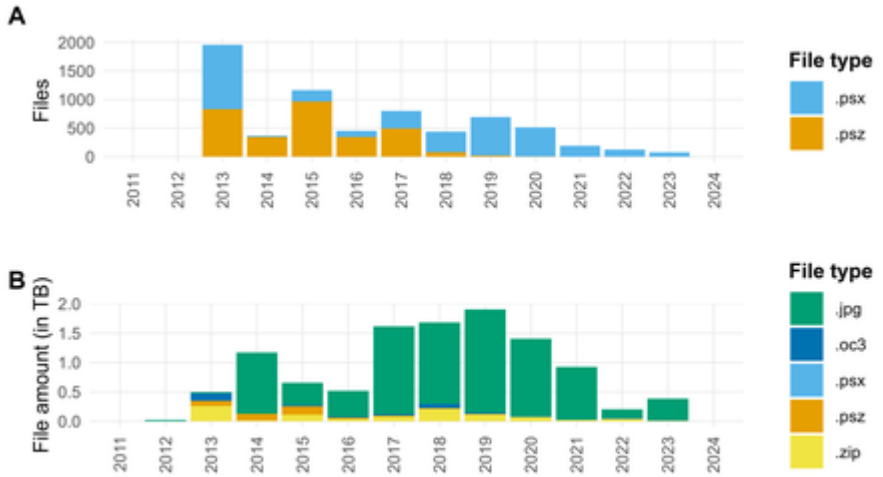

Figure 5: Temporal distribution of Metashape project files (A) and SfM raw data (B). Graphic: Jonas Abele.

In projects before the introduction of the guidelines, photographic documentation was often stored in JPG files, but since 2018 TIFF formats have been mandatory for photographic data. Nevertheless, a considerable number of JPG files remain (see the file list of Phase 4 above). A large share of these files can be attributed to the submission of JPGs collected to create SfM models. The marked increase in data volumes between 2018 and 2021 is due to formats generated through the SfM process. In addition, Metashape project files were frequently submitted (Figure 5A), illustrating how Metashape as a software solution has now become firmly established across the workflow.

From an archival perspective, these raw data and project files present several problems. More than one terabyte of storage is now occupied solely by the proprietary and non-preservation-ready Metashape project files. Identifying the actual raw data proves more difficult, since the JPGs recorded for SfM models cannot be distinguished from other JPG files based on file format alone. Moreover, the project data in the .psx format are structured as ZIP containers, which similarly cannot be automatically classified as SfM raw data solely based on their file extension. In many cases, however, the data structure follows a largely standardised folder naming convention with the suffix SfM, which made it possible to construct a vector to list all potential SfM raw data within such SFM-labelled folders and thereby obtain a reliable, albeit not entirely precise, impression of the distribution of SfM raw data in the long-term repository (Figure 5B).

3.4 Conclusion: data analysis

A comprehensive data analysis of long-term storage was systematically conducted for the first time in 2023 and has since become an essential component of the preservation strategy at LAD-BW. On the one hand, the data analysis serves as data evaluation and monitoring; on the other hand, and as a priority, it provides the basis for identifying specific preservation requirements.

The data collection, which has grown over the decades, is highly heterogeneous regarding data structures and file formats. Since 2018, the guidelines have regulated the selection of preservation-worthy information for the majority of newly generated primary documentation data. The file format specifications ensure preservation-ready data, and through defined metadata and standardised data structures, the adherence to the FAIR (Findable, Accessible, Interoperable and Reusable) principles has been significantly improved. In practice, however, ensuring that data packages comply with the guidelines proves challenging. The large number and size of projects make corrective intervention during ingestion difficult, not least due to the limited personnel resources available for preservation, which usually focuses on cleaning up the ADAB folder.

Particular difficulties arise with data bundles generated using new methods and collected on a large scale, if these methods are not yet covered by the guidelines and requirements, or if they are not covered precisely enough. In long-term data management, such problems often only become apparent with a certain time lag, as several months may elapse between the collection and the submission of data. A particularly illustrative example is provided by the SfM data. As a method, SfM has become firmly established, especially for the large areas of rescue excavations. The possibility of rapid data capture by drone has effectively become a de facto standard, as rescue excavations are conducted under significant time pressure.

4. Legacy data project

Within NFDI4Objects, LAD-BW is jointly responsible for developing standards and guidelines for the creation and data management of primary documentation data. One aspect of this task is the development of data management plans for the digital primary documentation data of archaeological heritage management in Baden-Württemberg. Another key focus is the strategy for dealing with digital legacy data. Such data are currently the subject of a dedicated working group, which has developed a strategy already being implemented in practice. Initially, the emphasis is on data hygiene, which under certain circumstances includes the deletion of data. In addition, preservation planning and data migration are central issues, specifically concerning the long-term preservation of digital data and the sustainable and efficient use of resources.

Deciding what is not worthy of preservation is not straightforward. The difficulty partly stems from the fact that the data submitted to the excavation data storage were deliberately compiled by the respective project staff as complete primary documentation packages, with the explicit aim of long-term safeguarding and in the expectation that they would remain unaltered.

Data analysis, however, has found that conducting data hygiene under clearly defined criteria is not only appropriate but also indispensable. A prominent example is the SfM data from excavations, which can account for a considerable share of the total data volume of excavation documentation. As shown in the analysis of storage consumption (see section 3.3), particularly in the years 2018 to 2021, large quantities of SfM data were deposited, including raw data (photographs; Figure 5B), processed Metashape project files (Figure 5A), and DEMs and orthophotos exported from the models. Many of these datasets are

associated with non-preservation-ready formats and redundant information, highlighting the importance of systematic data hygiene.

4.1 Selection and preservation planning

Given the vast amount of digital data that inevitably arises in the digital age, selection is indispensable (DPC 2015, 12). In the data packages of individual projects, we repeatedly encounter cases in which far more data were submitted than would have been required from an archival perspective. Particularly in the legacy data projects, there are many files not worthy of preservation, such as intermediate backups — created no doubt with the best of intentions but never removed from the SIP — or photo intake folders in which daily photographs from the camera were initially stored, subsequently integrated into the folder structure, and renamed but not deleted from the intake folder. During ingestion, such data were often accepted without further selection, given the limited resources available. For these projects, however, it is not only justified but also, to ensure reusability, essential to conduct retrospective selection and disposal under clearly defined rules.

Given the large volume of digital legacy data and non-preservation-ready holdings in the long-term storage of primary documentation, migration — in addition to selection and data hygiene — is a central issue. In line with the OAIS reference model, digital migration is understood as the transfer of digital information for preservation, characterised by a focus on maintaining the full information content, the complete replacement of the old implementation by the new, and that the OAIS retains full control over all aspects of the transfer (Consultative Committee for Space Data Systems 2024, 5-2).

As both areas (selection/deletion and migration) represent a profound intervention in the data, careful documentation of the processes is indispensable. Identifying the need for action and each individual step of the migration must be recorded in a transparent and traceable manner (nestor-Arbeitsgruppe Dokumentation in der digitalen Langzeitarchivierung 2025, 35). File deletion, too, may only be conducted in accordance with an admissible strategy, as specified in OAIS (Consultative Committee for Space Data Systems 2024, 3-1).

In the course of the legacy data curation workflows, all processes have thus been systematically documented, ensuring that the selection and migration decisions taken remain transparent and permanently comprehensible. Therefore, standardised processes have been established and fixed workflows defined, which are outlined in the following section.

4.2 Documentation strategy: standardised processes

For all work involving the digital primary documentation data, clearly defined, standardised processes are established. These processes link decision points with specific actions and set out the basic prerequisites for carrying out particular migration or selection steps. They thus provide a schematic framework within which concrete workflows are defined for practical implementation. Both the applied process and the individual steps carried out for a data package are documented in tabular form and appended to the data package. A mandatory component of this documentation is the creation of file lists including SHA256 checksums (both for the original state of the files and after completion of the standardised process). The file lists and the related documentation are stored within the SIP in an additional folder, thereby ensuring the traceability of all changes made.

The standardised processes follow a clear naming convention and are introduced whenever a migration or deletion step, such as the conversion of RAW images, can be performed repetitively and in a largely standardised manner across multiple projects. Each process

carries the prefix 'sp' for 'standardised process', followed by an underscore and an abbreviation for the respective subject area. At present, the defined areas are DEL (Delete) and DM (Data Migration). This is followed by a four-digit number, starting with 0001, which uniquely identifies the process and must not be altered. If a process is modified or extended, this is indicated by a version suffix (e.g. v1, v2) appended to the process name. While the fundamental objective of the process remains unchanged, the process may be adapted to meet new requirements.

Standardised processes can be further developed, whereby earlier versions are not necessarily replaced but may continue to be applied if they prove suitable in the context of other data packages. In practice, the documentation of these data curation steps, such as migration, transformation and controlled deletion of data, through standardised processes has proven to be an effective means of ensuring that deletion procedures (particularly those considered critical from an archival perspective) are described transparently and remain fully verifiable.

4.3 Standardised process for the curation of SfM Raw Data 'sp_DEL_0001_v1'

As the first case study of a standardised process, 'sp_DEL_0001_v1' is outlined (Figure 6). This process schematically describes the procedure for deleting non-preservation-ready Metashape project files, provided that various parameters are met.

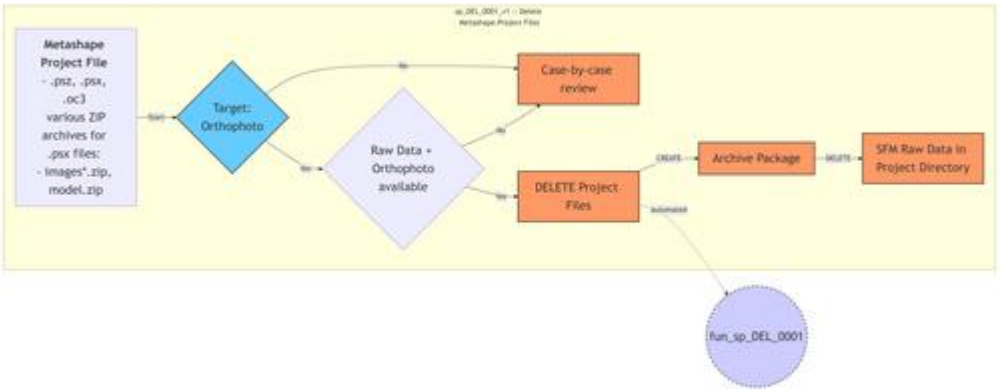

Figure 6: The standardised process 'sp_DEL_0001_v1' for documenting the deletion of Metashape project files. Graphic: Jonas Abele.

In most cases, the SfM documentations found within the data packages of excavations and trial trenching — that is, the ground-intervening methods — were conducted to produce an orthophoto, a planum, or a profile.

The primary objective, the generation of an orthoimage, fulfils the first parameter (Figure 6, blue diamond) and must be fulfilled for the process to apply at all. In contrast, if the three-dimensional (3D) documentation concerns a complex archaeological context, such as a deposit feature, whose primary objective is not the production of an orthophoto, then an

individual case assessment is required. In such cases, extensive postprocessing steps — for example, cleaning the point cloud of artefacts or mesh optimisation — are often performed within Metashape and are thus stored in the project file. Moreover, the end product is not a two-dimensional (2D) orthophoto but a 3D file that needs to be checked for existence and suitability. The archiving of complex 3D scenes prepared in Metashape project files could certainly be incorporated into a future process but is not the current focus, as the majority of SfM data serve documenting plana and profiles to generate orthophotos.

If the primary purpose of the Metashape project file is orthophoto generation, the standardised process further checks whether the raw data in the form of photographs (.jpg) and measurement data are available. The archiving of raw data ensures the SfM model can, in principle, be reproduced. Another part of this decision node (Figure 6) is that the georeferenced orthophoto must be available as a GeoTIFF. If the raw data or the orthophoto are missing or only partially present, an individual case assessment is necessary.

It should be emphasised that the raw data of these SfM files continue to occupy a very large volume of storage space. As the desired result of an orthophoto already exists, they rarely need to be disseminated as part of a DIP; they only become relevant if the 3D models are to be recalculated. Therefore, in a final step before deleting the Metashape project file, the raw data are transferred into a distinct data package and separated from the other primary documentation data at the project level within the directory structure.

This strategy is intended to be incorporated into future excavation guidelines, whereby the SfM raw data are submitted as an independent data package, focusing on the spatial situation and the moment of recording rather than, as previously, on the documentary, archaeologically defined result of a planum or profile. The package will include and be archived as a compilation of the following components: all raw data collected for the model (photographs, JPGs), the coordinates of the reference points used as a plain-text file, a metadata file, a location plan, and the model itself (glTF/GLB format). The glTF (GL Transmission Format) is an open, fully documented International Organisation for Standardisation (ISO) standard developed by the Khronos Group and widely supported across 3D applications and web environments. The contexts, which may contain several horizontally or vertically definable archaeological surfaces, will be recorded and listed as packages with individual numbering.

It is assumed that the aim of data collection is defined prior to documentation. This consideration determines both the recording method and the mode of storage and ultimately decides whether the 3D data are transferred to the storage system. The underlying principle is that models created to capture a spatial situation can be regarded as preservation-worthy (primary models). Mandatory documentation of the 3D context is required, for example, for burials, deposits or architectural structures (including architectural remains).

In contrast, models created merely for the extraction of orthophotos — such as those documenting archaeological plana and profiles — are defined as secondary models. These models are not considered preservation-worthy and are not to be submitted to the LAD as part of the primary documentation. Only the resulting raster images are requested, in accordance with the guidelines, and can be easily included into the canon of defined documentation components.

A detailed description of this 3D documentation package for archaeological primary documentation, including specific selection criteria and mechanisms for quality control, is in preparation.

4.4 Duplicate detection and removal: 'sp_DEL_0002_v1'

Duplicates also fall into the category of files to be deleted. In the data analysis, a particularly large potential for action was identified in the area of image data, which is why the process described here focuses on image file duplicates — that is, files with identical or nearly identical image content stored multiple times. Such duplicates may arise in several ways:

- Multiple storage in different directories: files are often stored in several folders organised according to different criteria (e.g. by features), resulting in separated copies of the same image material.
- Format-based duplicates: the same image content may exist in different file formats (e.g. JPG, TIFF, BMP, PNG) created for different purposes.
- Minimally edited variants: minor adjustments to an image (e.g. brightness corrections, cropping) lead to files that are nearly identical in content but exist technically as separate files.
- Backup copies and safeguarding strategies.

This uncontrolled proliferation of file copies is a major problem in data packages created before the introduction of the excavation guidelines in 2018, and necessitates systematic cleaning to consolidate the data holdings.

To identify duplicates, SHA256 checksums are first calculated, which reliably detect exact duplicates. For files with different formats or slight modifications, content-based comparison methods may be applied. One option is perceptual hashing (using the Python library *imagehash*), which generates image signatures that are robust against minor changes (e.g. colour corrections, noise). By comparing the Hamming distance between hash values, similar images can be identified. Another option is feature-matching (using OpenCV), in which visual features (e.g. edges, textures, key points) are extracted and compared to detect matches even after more substantial edits or format conversions. Both approaches have so far been discussed primarily on a theoretical level and require practical validation. Although these approaches enable the detection of similar files content-wise, they provide no absolute certainty but only probabilities of similarity. A final manual review is therefore indispensable.

Not every identified duplicate is automatically subject to deletion. The decision regarding which files are retained and which are deleted depends on the filing logic:

- Duplicates in the source directory are removed once the converted files are present in the target directory.
- Duplicates in the target directory are retained if they are contextually justified, for example, if a photo depicts several features stored in separate folders.

Backup folders are gradually dissolved, so that ultimately only one consolidated version of the data remains.

It is essential that the original file names are preserved to ensure that references (e.g. in photo registers or excavation documentation) remain valid.

4.5 Migration of RAW image data: 'sp_MG_0001_v1'

The guidelines prescribe TIFF as the archival format for photographs. Nevertheless, RAW files from various manufacturers continue to be regularly deposited, even in recent projects. These files occur particularly frequently in older projects (Figure 7), so that, within the

framework of the Legacy Data Project and in the context of a standardised process, an appropriate strategy needed to be developed. RAW image files represent the direct, unprocessed output of the camera sensor and contain extensive metadata in the Exif data, including technical information, such as camera model, exposure settings, capture date, and lens data. As these files are not post-processed, they are regarded as digital masters. However, most RAW formats are proprietary and not openly documented.

The standardised process 'sp_'MG_0001_v1' is quite straightforward, as its aim is to generate an archival TIFF file from a RAW file that complies with the specific technical and formal requirements of the archive (see below). Nevertheless, both from a data management perspective and regarding the technical implementation, several challenges have arisen, which are discussed in more detail in the following section.

Figure 7: Distribution of RAW image data in long-term data storage. Graphic: Jonas Abele.

In many cases, RAW files already exist in converted form (as JPG or TIFF) and could, therefore, in theory, be classified as photo duplicates (see above). However, this situation offers an opportunity for quality assurance. If the already developed images do not meet the applicable archival standards e.g. regarding colour depth or compression, the raw data can be reconverted, this time based on the current specifications.

4.5.1 Digression: TIFF as the standard format for long-term preservation

To minimise the risks associated with proprietary formats, LAD-BW has adopted TIFF as the standard format for archiving. TIFF is an openly documented, licence-free format characterised by high compatibility and lossless storage. The format is also recommended by, among others, the German National Library for long-term preservation (International Telecommunication Union [ITU] 1992; Deutsche Nationalbibliothek [DNB] n.d.; nestor 2010; Langzeitarchivierung.de 2024).

The TIFF files intended for submission to the long-term storage of primary documentation data at LAD-BW are generally based on the Baseline TIFF 6.0 specification and must meet the following requirements:

- Uncompressed or with lossless LZW compression to preserve maximum image quality. Although converting RAW to TIFF generally results in larger files, this format guarantees that as little information as possible is lost. In the case of uncompressed TIFFs, even if bit errors occur — for instance, due to defective storage media —

parts of the image often remain readable, whereas compressed formats, such as LZW-TIFFs, may render entire blocks of the image unreadable when errors occur.

- An ideal colour depth of 16 bits per channel (or 16-bit greyscale) ensures optimal preservation of dynamic tonal ranges and subtle colour gradations. This depth corresponds to the recommendations of the nestor handbook and the DFG's practical guidelines for digital long-term preservation (nestor 2010; Deutsche Forschungsgemeinschaft 2016). As a minimum standard, eight bits per channel should not be undercut, in order to guarantee sufficient image quality. In principle, colour depth should correspond to the maximum output capability of the camera but must not fall below the specified minimum.
- AdobeRGB (1998) colour space, which offers greater colour fidelity and a wider gamut than sRGB. Alternatively, ECIRGB_v2 or ProPhoto RGB may be used, provided compatibility with downstream systems is ensured.
- Embedded ICC profile, which guarantees that colours are interpreted correctly regardless of the software used.
- Comprehensive metadata integration, including Exif data (technical camera settings), XMP data (e.g. title, author, descriptions), and METSRights (rights information). Optionally, metadata may also be stored in a separate XML sidecar file to ensure redundancy.

4.5.2 Challenges in conversion

A central problem in converting RAW to TIFF is the potential loss of metadata. Many standard tools (such as IrfanView) do not convert metadata completely into the target format. Although IrfanView preserved metadata during the conversion from RAW to JPEG, it did not do so when converting to TIFF, which led to information loss that is unacceptable from an archival perspective.

Therefore, LAD-BW relies on a combination of RawTherapee and ExifTool:

- RawTherapee is employed due to its neutral development mode, which converts raw data into TIFF with minimal software-induced alterations. This process ensures the content of the TIFF file remains as close as possible to the original sensor signal.
- ExifTool serves as a control mechanism to guarantee that all available metadata from the RAW file are transferred into the TIFF file. If gaps are identified, ExifTool can subsequently be used to fill them.

4.5.3 Workflow for processing RAW image data

As part of current data hygiene measures, a systematic analysis of directory structures is conducted to identify and eliminate duplicates. In this process, RAW files — once converted into preservation-ready formats (TIFF or JPG) — are deleted to avoid redundancy and consolidate the data holdings. This process ensures that only a single, high-quality version of each file is retained, and the original RAW files are removed as unnecessary duplicates.

The conversion of raw data is performed via Python scripts that integrate the open-source tools RawTherapee and ExifTool. Well-tested scripts, which have proven their reliability in practice, are intended in future to be published on platforms such as GitHub, making them accessible to the professional community.

After successful conversion, the new TIFF files are renamed according to internal naming conventions and moved into the designated target directories. The original RAW files can then be deleted, since the TIFFs now exist in sufficient quality and serve as preservation-ready masters.

4.5.4 Handling of DNG files

For the time being, DNG files are retained in their original form within the LAD storage. Although DNG is considered an open standard and is documented in detail in the Adobe DNG specification, the long-term future of the format is not entirely secure (Adobe Systems 2024). As long as DNG is regarded as an open standard and the overall data volume remains manageable, LAD refrains from any preventive conversion to TIFF. However, should Adobe restrict the specification in the future, subsequent conversion is foreseen.

4.5.5 Headerless .raw files

Five data packages of excavation results contained headerless .raw files whose manufacturer and specifications could not initially be identified. Common tools, such as IrfanView, Darktable or ExifTool, failed to decode these files. Further analysis methods, including RawTherapee, DCRAW and hex viewers, were tested but also proved unsuccessful. It could only be determined that these were headerless .raw files. Fortunately, these files already existed in converted form as TIFF and JPG, though not of sufficient quality.

As a result, a script was developed that incorporates known camera specifications (which could be extracted from the converted JPG files) and performs a conversion from .RAW to .TIFF based on file size and the expected pixel ratio. In addition, a heuristic conversion option was implemented that does not rely on camera specifications. This method attempts conversion by considering file size, expected aspect ratios and pixel distribution, but this is an initial trial-and-error approach. Ideally, however, the camera and its specifications should be known to ensure a higher probability of successful conversion.

5. Conclusion and future directions

The heterogeneous and historically evolved collection of archaeological primary documentation data in the long-term storage of the LAD-BW represents a significant challenge, particularly regarding preservation planning. This challenge can only be addressed gradually and through sustainable strategies. The Legacy Data Project, initiated within the framework of NFDI4Objects, provides an opportunity to address this complex issue systematically.

The developed standardised processes and associated workflows have been successfully tested in several data packages within the long-term storage of primary documentation data. Standardised TIFF conversions, for example, enable the transformation of proprietary formats into open, preservation-ready standards. The targeted use of RawTherapee for neutral raw data development and ExifTool for comprehensive metadata transfer ensures lossless conversion, and project-based automated duplicate detection using SHA256 hashes combined with manual plausibility checks reduces redundancy and frees up storage space.

It is already evident that migration processes, especially for digital legacy data, cannot be limited to file formats alone. There are projects, for instance, in which photos (often duplicates) were organised within folder structures according to archaeological entities with no accompanying metadata, such as a photo list, so the folder structure itself functions as a

carrier of information. In legacy data in particular, we frequently encounter project situations that reflect highly heterogeneous practices and often bear the individual imprint of excavation technicians. These situations lead to information interdependencies between different data objects that may not necessarily exist in interoperable environments, such as GIS or databases, but can only be reconstructed indirectly through folder names, file names or readme files. Such constellations reveal that, even in the absence of binding standards or guidelines, “data chaos” does not inevitably result; rather, we often find carefully considered but rarely documented solutions. Preservation planning for such projects is therefore always a hermeneutic process that must go beyond the level of individual files.

In addition to continuing the Legacy Data Project, two central fields of action are emerging for LAD-BW that will further improve primary documentation data: first, the revision of excavation guidelines; and second, the development of new target systems — including a geodatabase for excavation-related spatial data — that will enable the user-friendly creation and management of guideline-compliant primary documentation data.

In addition to these developments, the associated improvements in data management and data quality — resulting in the creation of preservation-ready primary documentation data — also foster the establishment of systemic workflows and standards. These form the basis for strengthening the FAIR principles and for ensuring their long-term implementation.

A central aspect in advancing the FAIR principles is the continued development of standardised and centralised metadata within the new target systems, together with a more structured and interconnected knowledge base. These measures are expected to contribute to making primary documentation data more findable through consistent identifiers and metadata, more accessible through harmonised storage and publication workflows, more interoperable through the use of open standards, and more reusable through improved contextualisation and documentation. Furthermore, two open-access publication series have been established as part of the 'Pilotprojekt In-wertsetzung Ausgrabung (PIA)' conducted by the LAD-BW, with the aim of making the results of archaeological investigations publicly available as soon as possible after the completion of fieldwork (Krausse *et al*. 2024, 11–18). Within the series 'Dokumente zur Archäologie in Baden-Württemberg', 15 reports have already been released, while the more extensive series 'Materialien zur Archäologie in Baden-Württemberg' currently comprises two volumes, which already include research data publications as integral components.

As noted by Bibby (2021), *'the challenges of the coming years'* — transforming the extensive data pool of the LAD-BW and establishing sustainable structures for a modern archive aligned with the FAIR principles — remain relevant today. The current efforts continue along this path, though many obstacles still lie ahead.

Bibliography

ADABWeb n.d.'Was ist ADABweb?', [online] Landesamt für Denkmalpflege https://www.denkmalpflege-bw.de/denkmale/datenbanken/adabweb (Last accessed: 10 September 2025).

Adobe Systems 2024 'Digital Negative (DNG)', [online] Adobe. https://helpx.adobe.com/de/camera-raw/digital-negative.html (Last accessed: 9 September 2025).

Archaeology Data Service n.d. 'File formats for archiving datasets', [online] Archaeology Data Service. https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/data-collection-and-fieldwork/laser-scanning-for-archaeology/archiving-laser-scan-data/file-formats-for-archiving-datasets/ (Last accessed: 26 October 2025).

Bibby, D 2021 'Digital Archaeological Archiving in Baden-Württemberg, Germany: an evolving system', *Internet Archaeology* **58**. https://doi.org/10.11141/ia.58.3

Consultative Committee on Space Data Systems 2024 'Reference Model for an Open Archival Information System (OAIS)', Washington, DC: CCSDS Secretariat. https://ccsds.org/wp-content/uploads/gravity_forms/5-448e85c647331d9cbaf66c096458bdd5/2025/01//650x0m3.pdf (Last accessed: 15 September 2025).

Deutsche Forschungsgemeinschaft 2016 'Praxisregeln “Digitalisierung”', [online] DFG https://www.dfg.de/de/foerderung/foerdermoeglichkeiten/programme/infrastruktur/lis/veroeffentlichungen (Last accessed: 9 September 2025).

Deutsche Nationalbibliothek (DNB) n.d. '06 - Das Dateiformat TIFF'. https://files.dnb.de/nestor/kurzartikel/thema_06-TIFF.pdf (Last accessed: 9 September 2025).

Digital Preservation Coalition (DPC) 2015 *Digital Preservation Handbook*, 2nd edn, Aldus Corporation, Washington. https://www.dpconline.org/docman/digital-preservation/handbook/1552-dp-handbook-digital-preservation-briefing/file (Last accessed: 10 September 2025).

Firmenarchäologie n.d. 'Firmenarchäologie'. https://www.denkmalpflege-bw.de/geschichte-auftrag-struktur/archaeologische-denkmalpflege/firmenarchaeologie (Last accessed: 10 September 2025).

IANUS LZA-Empfehlungen n.d. 'Empfehlungen zur Langzeitarchivierung für die NFDI4Objects-Community'. https://ianus-fdz.de/it-empfehlungen/ (Last accessed: 26 October 2025).

International Telecommunication Union (ITU) 1992 *TIFF Revision 6.0* . https://www.itu.int/itudoc/itu-t/com16/tiff-fx/docs/tiff6.pdf (Last accessed: 9 September 2025).

Krausse, D., Ebinger, N. and Link, T. 2024 'Das 'Pilotprojekt Inwertsetzung Ausgrabungen' (PIA)' in D. Krausse, N. Ebinger and T. Link (eds) *PIA 1. Bericht des Pilotprojekts Inwertsetzung Ausgrabungen*, Heidelberg: Propylaeum (Materialien zur Archäologie in Baden-Württemberg 1), S. 11–19. https://doi.org/10.11588/propylaeum.1493.c21625

nestor 2010 'nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung', Version 2.3. https://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf (Last accessed: 9 September 2025).

nestor-Arbeitsgruppe Dokumentation in der digitalen Langzeitarchivierung 2025 'Leitfaden zur Dokumentation in der digitalen Langzeitarchivierung', Version 1.0, Frankfurt am Main: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen in Deutschland e.V. https://nbn-resolving.de/urn:nbn:de:0008-2507101336321.667080583730 (Last accessed: 13 September 2025).

Open Preservation Foundation 2022 'New Community Resource: International Comparison of Recommended File Formats'. https://openpreservation.org/news/new-community-

resource-international-comparison-of-recommended-file-formats/ (Last accessed: 26 October 2025).

Swedish National Data Service n.d. 'Recommended file formats'. https://snd.se/en/research-data-support/manage-data-descriptions-doris/recommended-file-formats (Last accessed: 26 October 2025).

VLAK-AIS (Kommission Archäologie und Informationssysteme des Verbandes der Landesarchäologien Deutschlands) 2024 'Archivierung – Basisanforderungen digitaler Archivierung in der Bodendenkmalpflege'. https://www.landesarchaeologien.de/fileadmin/mediamanager/004-Kommissionen/Archaeologie-und-Informationssysteme/Archivierung/Archivierung_Basisanforderungen_v1-00.pdf (Last accessed: 10 September 2025).

Wikipedia 2025 'Lotus 1-2-3'. https://en.wikipedia.org/wiki/Lotus_1-2-3 (Last accessed: 10 September 2025).