https://doi.org/10.2352/issn.2168-3204.2023.20.1.36


# Digitization Information Objects: The Case for Standardized Digitization Project Specifications

*Jim Studnicki; Creekside Digital; Glen Arm, Maryland, United States of America*

## Abstract

*While we now have mature, proven guidelines (FADGI) which provide solid recommendations on how to create proper master files, beyond targets and the ability to measure them, the cultural heritage community lacks easily consumable, flexible specifications for conducting actual projects. Moreover, there is a general lack of examples of FADGI-compliant Statements of Work, leading to much re-invention of the wheel and even to library and archival personnel deciding to not use FADGI at all. This puts inexperienced users at a decided disadvantage and creates a formidable barrier to entry for new practitioners who want to use the FADGI guidelines on their projects. As discussed in this paper, a DIO (or Digitization Information Object) is a data model encompassing all technical parameters of a still image digitization project. At its core, the DIO schema is intrinsically tied to FADGI, and enforces FADGI compliance through its use. It provides a common, machine-readable instruction set for digitization-facing software programs. This allows consuming applications to be quickly and precisely configured per-project to specify output image parameters, configure post-processing workflows, verify both working files and huge batches of completed content at scale, and even to provide plain-English text for a project's Statement of Work -- all from the same DIO JSON file.*

## The Problem with (No) Specs

No one builds a skyscraper without detailed engineering drawings – yet, again and again, that is how we see digitization projects continually conducted, representing a massive risk of money and effort. It has been 13 years since the First Edition of the FADGI guidelines was released. While FADGI has become wildly successful when its use has been mandated, the author estimates that over 80% of the digitization RFPs and Request For Quotes that come across his desk at the time of this writing either fail to mention FADGI entirely, or do so in such a way that is completely non-actionable (e.g., failing to require daily submission of targets from each scanning station, not explicitly requiring rework when a failing target is submitted with a batch, etc.). The cultural heritage community has free access to vetted guidelines which are backed by world-class image science and have been continually proven on a massive scale for over a decade, yet even inside our own small universe of projects, these critical guidelines see actual use only on a relatively small minority of digitization projects. Why is this?

## Increasing FADGI Adoptance through Project Standardization

Even though much progress has been made with the definition of digitization specifications (specifically since the advent of the FADGI guidelines), the cultural heritage community still suffers from multiple issues related to an overall lack of consistency in how these specialized efforts are defined and conducted. There is tremendous variability in project specifications from institution to institution, and even within institutions, project to project. Part of this is because of the intrinsic complexity of digitization efforts themselves, which are inherently technical activities requiring skills in both the photographic arts as well as knowledge in information technology. Other factors include the expense of the necessary equipment (again, both photographic as well as computing), the experience levels of the participants, and the need to support multiple and / or different uses cases with the resulting assets. Finally, and most critically, there is a lack of useful examples of digitization project Statements of Work which can be used as solid starting points. At a system level, one can think of all of these items as multiple variables – a huge array of levers, knobs, and buttons which, without guidance and standardization (and defaults), are continually and infinitely adjustable, constantly conspiring against the delivery of accurate and consistent results. In absence of tightly defined, repeatable controls over this landscape of entropy, the inevitable result is an ever-changing kitchen sink of requirements.

The key to eliminating much of this variability is to standardize the creation and enforcement of detailed project specifications which precisely define as much of the digitization project's technical requirements as possible. Moreover, these specifications should be locked to the FADGI guidelines (and, where applicable, the technical requirements of M-19-21 for permanent record digitization) in order to leverage the important work already built into those efforts, as well as to codify compliance with these guidelines and regulations into the resulting project specifications. Finally, to facilitate reuse and implementation, these specifications must be machine-readable (that is, easily ingested and interchanged by computer programs). This allows a multitude of technical information to be quickly consumed by the software and hardware components required for digitization activities, leading to a reduction in direct human interaction with the equipment and

workflow tools during production. This leads to fewer human-introduced errors and opens the door towards full workflow automation of most production activities. As a result, flaws and errors are identified and driven out of each batch of assets much earlier in the workflow, leading to higher quality content, fewer errors delivered to the customer, greater efficiency, and more value per dollar spent on core digitization activities.

## Components of the DIO Schema

The resulting schema as defined in JSON (JavaScript Object Notation) has been named Digitization Information Object, or DIO. It is a standardized data model that encapsulates all of the technical requirements for creating a batch of digitized still image assets. Its properties closely map to key attributes of the FADGI guidelines. For example, each DIO has one and only one Material Type. Its Material Type property maps directly to the categories found in the current (Third Edition) FADGI Guidelines as well as those used by the latest version of the OpenDICE software tool. Similarly, a DIO can have one and only one Star Tier (e.g., 1 Star, 2 Star, 3 Star, or 4 Star). Together, these two parameters set the default technical requirements for the master files in the batch as comprehensively defined in the FADGI Guidelines themselves. Typical parameters defining master files include file type, bit depth, number of color channels, compression type, ICC color profile, and horizontal and vertical pixel resolution. Master file parameters can be overridden as long as they meet the minimum acceptable values defined by FADGI. For example, for several Material Types, in practice it is common to increase the minimum resolution of 3 Star projects from 300ppi to 400ppi. Currently, this intelligence is enforced by the application used to create DIO JSON (e.g. Creekside Digital's DIO authoring software, called Composer); eventually, these defaults and allowable values may be codified into the JSON schema definition itself (i.e., similar to XSD validation for XML files) as many parts of DIO are simply implementations of the FADGI Guidelines rendered as JSON properties. In this way, the DIO schema will become more independent of the tools currently used to create it.

Along with the master files, a standards-compliant digitization project needs to define how it uses targets to measure the image quality metrics defined by FADGI. Each DIO allows the definition of a Target Type (e.g., ISA Device-Level, ISO 19264, etc.). Combined with the Material Type and the Star Tier, the definition of the Target Type in a DIO completes the information needed to support full automation of target analysis at a particular level of quality. Future updates to the DIO schema will include support of a "target manifest," which allows a single digitized object to consist of master files originating from different pieces of digitization equipment (e.g., a bound book whose pages are digitized with a camera on a copy stand, while its included fold-out maps were scanned separately on a large moving-table scanner, with separate target images originating from each platform necessarily contributing to the same book).

However, digitization project specifications necessarily go far beyond the technical master file guidelines and target measure-


```
{
  "id" : 362,
  "name" : "400ppi Bound Books",
  "standard" : "FADGI3",
  "material" : "BoundVolumesGeneralCollection",
  "materialTarget" : "Dice",
  "requireImageIndexPerTargetFile" : false,
  "checksumType" : "Simple",
  "hashingAlgorithm" : "MD5",
  "xmlBox" : null,
  "generalSpecifications" : [ {
    "id" : 367,
    "type" : "Master",
    "fileType" : "TIFF",
    "attributes" : [ {
      "id" : 368,
      "name" : "Extension",
      "value" : ".tif"
    }, {
      "id" : 369,
      "name" : "Resolution",
      "value" : "400"
    }, {
      "id" : 370,
      "name" : "UnitOfMeasure",
      "value" : "Inches"
    }, {
      "id" : 371,
      "name" : "BitDepth",
      "value" : "8-bit"
    }, {
      "id" : 372,
      "name" : "Color",
      "value" : "Color"
    }, {
      "id" : 373,
      "name" : "Colorspace",
      "value" : "Adobe RGB 1998"
    }, {
      "id" : 374,
      "name" : "Compression",
      "value" : "None"
    }, {
```


***Figure 1**. TIFF Master File Definition and other parameters as a snippet of DIO JSON. Note several other DIO parameters including Material Type ("material") and FADGI Star Tier ("standard").*

ments set forth by FADGI and M-19-21. Almost every digitization project creates at least one type of derivative file to support various use cases. These are often lightweight access or "reader" files, such as PDF versions of the master files enhanced with a searchable PDF layer, in both single- and multi-page versions. Other common derivative types include JPEG2000, JPEG, GIF, and non-image files including plaintext and XML.

These files may feed databases or drive downstream workflows, act as content on web-based presentation systems, or support any other use case imaginable where the typically huge, lossless master files are not needed or are undesirable. Multiple derivative types, along with all of their parameters, can be defined inside a single DIO instance.

Most digitization project specifications also define some image metadata. Indeed, while FADGI does not explicitly require image metadata to be collected, M-19-21 does, and sets forth a minimum set of technical image metadata fields to be correctly populated. Each DIO allows the selection of a metadata tag family (e.g., EXIF, IPTC, XMP, etc.) and then the definition of multiple tags in that family. The DIO can specify that any given tag is simply present and populated, or it can require that it is populated with a specific value or range of values. It is anticipated that the metadata functionality will be enhanced in future versions of the DIO schema to support additional tag families and more dynamic rules around metadata population. Sidecar metadata in a CSV file will be supported as well.

An intrinsic part of any digitization project specification are rules around how the files themselves are named and foldered. This is another aspect of digitization that is completely missing from both FAGDI and M-19-21, yet is often incredibly important in practice, as many times the names and folder structure of the assets themselves impart additional meaning and context. For example, the naming and foldering of a batch of digitized records may emulate the Box-Folder-Item organization of the real-world source records materials. Other naming and foldering implementations may represent the organization of documents in a multi-volume series by using volumes, parts, and sections. Yet other folder name schemes may represent issue dates, editions, and even sections of digitized newspapers. Numeric sequences may also be used to increment image and folder names.

In addition to allowing complete control over how master files and derivatives are named and foldered, the DIO schema includes the ability to define "tokens" – variables to which a character mask may be applied and then used multiple times throughout a digitization project. For example, a token called [ISSUE_DATE] might be defined by the mask YYYY-MM-DD representing a machine-sortable date as hyphenated 4 digit year, 2 digit month, and 2 digit day. This format is commonly used for naming folders containing a single digitized newspaper issue as well as a multipage PDF derivative representing the same issue as [ISSUE_DATE].pdf. In conjunction with the other tools mentioned above, tokens allow for the definition of almost any conceivable naming and foldering schema.

Finally, checksumming is another critical aspect of in-practice still image digitization that is not directly required by FADGI (though, again, it is mandated by M-19-21, though the specific implementation of checksums for any given project are left up to each Federal agency). The DIO schema allows for the support of a simple checksum manifest – a text file containing a list of all of the assets and the checksum of each. In this case, DIO also allows for the selection of the hashing algorithm used in the manifest, with

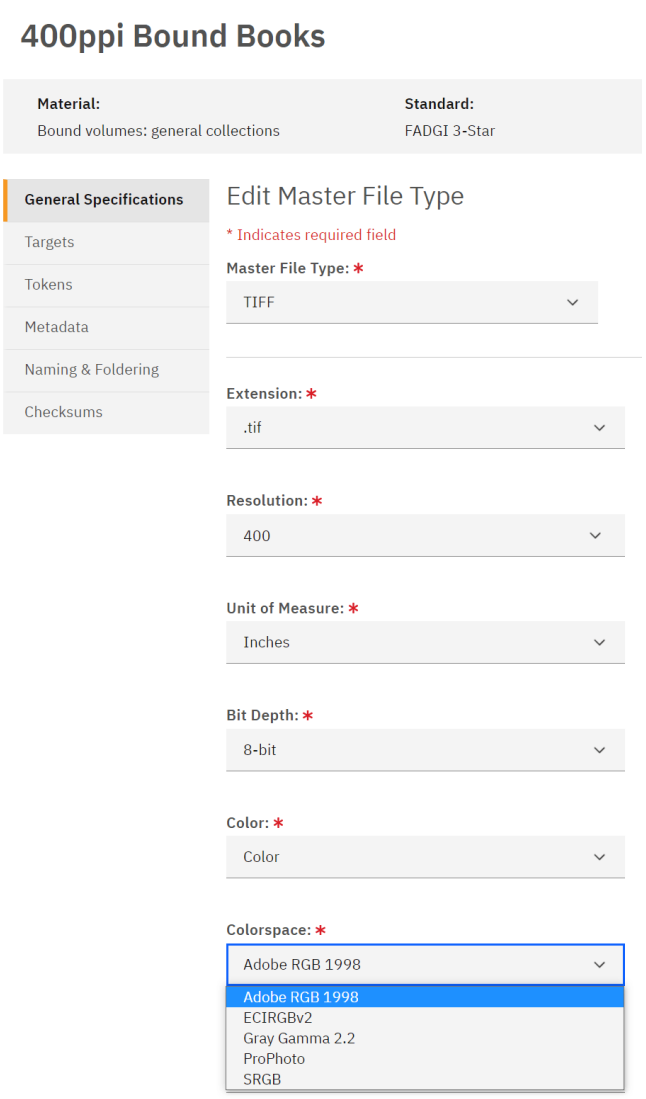


***Figure 2**. Human-readable, editable version of the previous TIFF Specification as rendered by Creekside Digital's web-based Composer application. The main categories of DIO parameters are generally represented in the navigation panel at left.*

MD5 as the default. Alternately, DIO allows the specification to elect the use of the Library of Congress' BagIt file packaging specifications, which contain integrated checksumming. Bag metadata can also be defined as part of a DIO instance. Note that use of bagging necessarily implies the use of a containing data folder and several supporting files, which has certain implications on project naming and foldering.

Once fully populated, a single DIO JSON file represents a comprehensive, portable, machine-readable digitization project specification. It can be transmitted as simply as emailing it to a colleague or vendor, and supports a complete universe of use cases regarding the creation and verification of digitization still image assets and the conducting of projects. For example, when

combined with completed batches of assets, a DIO JSON file could be used to drive massively scalable FADGI and M-19-21 compliance audits. The same DIO could reconfigure a software-based digitization workflow platform used to create finished assets from raw scans, ensuring that the proper types of master files were created, collecting all of the required pieces of metadata, generating the defined derivatives from QA-complete master files, and then naming and foldering and checksumming all of this content into a completed batch. Eventually, the same JSON file could even be consumed by DIO-compliant hardware platforms (e.g., scanners and reprographic camera systems) to change hardware settings, adjust camera height and aperture / shutter speed, illumination parameters, and output file settings, as well as to continually verify images coming off of the capture device to catch potential errors further upstream, up to and including while the source materials are still on / in the capture device for the first time. Other programs (such as Creekside Digital's Composer application) read and write DIO JSON and can also output human-readable, plain-English Statements of Work. These will vary from snippets of text to comprehensive, highly detailed technical specifications for an entire, large-scale digitization effort. The end result is human-readable content that can be used throughout various parts of a digitization project: during grant applications to demonstrate compliance with guidelines and standards, through procurement of services (i.e., to send a project out to bid and select a vendor), through execution of the project, and ultimately to verification and acceptance of the final digitized assets.

## Limitations

For various reasons, not every aspect of a digitization project is definable by the DIO schema at this time. Most of these gaps will be filled in incrementally in subsequent versions of the schema as feedback from users is received and as more resources for development become available.

One of the first items typically found in a digitization project's Statement of Work is a comprehensive description of the materials to be converted, along with their requisite quantities. It is especially important to note any particulars of material condition which might affect the digitization process. For example, in larger collections, different types and amounts of special handling may be required to completely and safely digitize a group of materials. Often, images of the actual collections are very helpful in illustrating to potential digitization practitioners what to expect. While DIO specifies the material type as defined by FADGI, it currently omits descriptions of the specifics of the collection, quantities, and categories of any potential special handling. We will explore adding these parameters in the next version of DIO, but in practice it may prove quite challenging to create a data model which supports the universe of all potential material specifics.

Descriptions of the materials, quantities, and types of special handling are also necessarily required to bill any digitization project. Even if the project is being conducted in-house by an institution's own personnel, this information is still critically important to ensure that the project is still progressing according to schedule and within budget. Often, pricing (or cost) is expressed as a series of line items based on material type, with an appropriate unit of measure (e.g., newspaper capture @ $0.75 / page). DIO omits pricing information and line items for now, but these will be important when soliciting bids from multiple vendors to ensure that the institution or agency is evaluating proposals using an apples-to-apples comparison.

Currently, the DIO schema does not allow for the definition of post-processing activities that manipulate the borders of individual page images. Principally, these activities include the splitting of 2-up captures / frames into individual lefthand and righthand page images, image cropping, and image deskewing. This is a glaring omission, as these are common activities defined in nearly every digitization project specification. Though it is certainly possible to include them in the schema today (and in fact, they will appear in the next version of the DIO schema), the first version of the specifications focuses only on parameters which can be conclusively verified by a machine. Verifying individual image crops and deskews for compliance with specifications such as "crop to within ¼" of the page edge while retaining all borders" and "no greater than 2% angle of skew of the text lines" is much more complicated (and currently not possible with software tools – only with human observation), and universal crop / deskew itself are not yet a 100% solved computing problem. We look to the application of machine learning combined with tens of millions of images of training data to solve problems like universal crop / deskew in the future, as well as the verification of these operations. The first step, of course, is their proper definition, and even though automated verification is not possible today, these parts of a specification still need to be included in any digitization project definition so that the desired output is precisely defined.

The digitization of three-dimensional objects is something also not formally addressed by the DIO schema. The FADGI guidelines deal principally with two-dimensional / classic "reprographic" digitization, and several FADGI metrics (e.g., Tone Scale and Uniformity) are often not appropriate when photographing objects on a traditional tabletop set. Moreover, DIO is not intended to define the parameters of three-dimensional data such as point clouds and or even the overlapping hemispheres of images commonly used to drive photogrammetry workflows. More input is needed from the museum community, and specifically from large museums which typically conduct mass digitization projects at scale.

Finally, as currently implemented, the DIO schema is inherently US-centric. This is by design as DIO has originated out of an effort to build tools and applications to enforce compliance with FADGI and M-19-21 for permanent records, which are of course guidelines and regulations created for and principally used by the United States Federal government. However, we see applications for this schema far beyond our own borders, and therefore seek feedback not only from digitization practitioners in the United States but from our colleagues around the globe. An incremental step would be the potential inclusion of the Metamorfoze Preservation Imaging Guidelines' levels of quality, which could be incorporated much in the same way as we have included the

various Star tiers of FADGI – though syncing Metamorfoze to DIO's material types may not be as straightforward as with FADGI (because of course, DIO's material types track FADGI directly). An additional area of exploration may likely center around Unit of Measure other than Inches, though on its surface this issue seems much easier to solve and should be at least partly supported by the current data model.

## The Future

Despite the above limitations, DIO is already proving itself useful today. The format drives Creekside Digital's first two software products, and we plan to open source the DIO JSON schema later this year (target date: Q4 2023) via an upcoming website. At the same time, we also plan to make our Composer application freely available to qualified institutional members of the cultural heritage and records management communities in order to assist them with defining their own standards / guidelines-compliant digitization project specifications. We view Composer as an expert system which will eliminate the need for human practitioners to memorize or continually consult and transcribe / interpret 120 pages of technical documents in order to produce and manage FADGI- and M-19-21-compliant specs. Most importantly, we will provide human-readable example Statements of Work on this website, and we will allow Composer users to create and export their own DIO objects and the accompanying standards-compliant Statements of Work from our Composer tool for use in their own RFPs / RFQs and resulting projects. As others in the cultural heritage community have noted, usage of a tool or a set of guidelines is a measure of success, and we are highly invested in increasing adoptance of the FADGI guidelines. We view the standardization of digitization project specifications themselves as the next step forward towards full industrialization as well as mainstreaming the use of these technologies, especially regarding the inclusion of non-expert users and practitioners. While today, FADGI-compliant projects are almost entirely conducted by large cultural institutions and university libraries, along with a small handful of highly specialized vendors, with the proper tools and training, along with mandatory regulations (e.g., M-19-21 for permanent records digitization in the United States), the performance of proper still image digitization in accordance with the guidelines can become mainstream and eventually the default.

## Author Biography

***Jim Studnicki is the founder and President of Creekside Digital, one of the largest standards-compliant service bureaus in the United States. Operating a 16,000 square foot facility just 55 miles from the U.S. Capitol, Creekside Digital processes some of the nation's most complex and challenging still image digitization projects. Jim holds an M.S. in Information Systems from the University of South Florida in Tampa.***