



Digital Succession Identifiers

2022-11-13

E. Castedo Ellerman  (castedo@castedo.com)

Abstract

DOCUMENT TYPE: Technical Specification

The Digital Succession Identifier (DSI) is a persistent identifier for bibliographic references. This specification of DSI is specific to the initial implementation using git.

Introduction

A non-technical background and motivation for this new technology can be found reading [Why try Digital Succession Identifiers](#) and in the [description of the perm.pub pilot project](#).

Digital successions

A digital succession contains digital objects. A digital object in this context is a fixed finite collection of bits. A computer file is an example of a digital object. A file system directory (folder) can also be represented as a digital object by both [git](#) [1] and [Software Heritage](#) [2]. Conceptually, each digital object in a digital succession is a new edition of a previous digital object. Each edition is assigned a number which determines its order. Although a digital succession expands, each digital object within the succession does not change.

A digital succession is also digitally signed. In the simplest case, a single user signs using a digital key that is verified to be for that user. In this specification, the digital signature is made using a [PGP key](#) [3]. When a digital succession is first created, it consists of only a digitally signed genesis record and no digital object editions. A *Digital Succession Identifier* is an intrinsic identifier [2] [4] of the genesis record.

Digital successions implemented via git

In the git implementation, the genesis record is a signed initial git commit with an empty tree (and no parent). To expand the digital succession, additional git commits are made using the same signing key.

The git tree of each commit is not a digital object in the succession. The git tree of a commit is a record of all the digital object editions in the succession. The top level directory consists of only subdirectories named as non-negative integers. Within each integer named subdirectory is either an entry named object or entries named as non-negative integers that are subdirectories. An entry named object is a digital object edition in the digital succession. This digital object can be a file (git blob) or a directory (git tree). For example, when a single file is added as edition 1, the full succession record is a directory with one path of 1/object to a git blob which is edition 1.

Multilevel edition numbering

In the simplest edition numbering scenario, edition numbers are just positive integers. For more advanced usage, multilevel edition numbering may be used. Multilevel numbering is found in the numbering of chapters, sections, and subsections (e.g. chapter 2, section 2.4, subsection 2.4.3). as well as software release versions (e.g. software release 2.19.2).

An edition number prefix, such as 1, can specify either a digital object edition or the entire sequence of editions 1.1, 1.2, 1.3, etc... An edition number specifies either a digital object edition or a sequence of subordinate edition numbers, but not both. Larger integers indicate newer editions that obsolete older editions with smaller integers. The DSI specification does not specify any semantic meaning to different number levels.

Textual representation of a Digital Succession Identifier

The textual representation of a Digital Succession Identifier (DSI) contains a base identifier followed by an optional slash followed by an optional edition number (possibly multilevel). The optional edition number is represented as non-negative integers separated by periods.

The base DSI is calculated from the git initial commit (genesis record) of a digital succession. Git-compatible software can calculate a 20 byte binary hash which identifies that digital succession. In most uses of git, this 20 byte binary hash is represented textually as a 40 digit hexadecimal number. In contrast, for a DSI, this 20 byte binary hash has a 27 character representation in standard base64url format (RFC 4648)[5].

It is worth noting, that when a new digital succession (git initial commit) is created, a user can choose not to use it and instead immediately create a new one with a different base64url text representation. There is little cost in not using empty digital successions and recreating new genesis records (git initial commits) until an acceptable base64url identifier is found.

Examples

Base DSI of this specification:

```
dsi:ji2STto1mZ3i2BmnGxbkebejKH4
```

DSI of the first edition:

```
dsi:ji2STto1mZ3i2BmnGxbkebejKH4/1
```

DSI of the first digital object (subedition) of the first edition:

```
dsi:ji2STto1mZ3i2BmnGxbkebejKH4/1.1
```

Future extensions

There are three paths to extending this textual representation to support future enhancements. They are to use a base DSI where

- some character is neither a slash (/) nor one of the 64 base64url characters,
- the number of characters is different than 27, or
- the 27-th character is one of the 48 base64url characters that never appear as the last character of a base64url encoding of 40 bytes (that is any base64url character that is not A, E, I, M, Q, U, Y, c, g, k, o, s, w, 0, 4, or 8).

Design Choice Rationales

Use of the term edition

The term edition is used to refer to a digital object in a digital succession. Part of the motivation is to avoid confusion with the many uses of the term version. For instance, version can be used to refer to versions of a digital object that are not added as editions.

This language is partly due to the initial application of digital successions for scientific articles. Although there are rarely new versions of scientific articles after they are published, it is well established that scientific textbooks are published in multiple *editions*.

In the initial application of digital successions, documents are the digital object. A single edition of a document might be presented in different HTML and PDF *versions*, even though they are the same *edition* of a document.

Separation of git history from edition history

A significant design choice is to not directly rely on git commit history to determine the succession of editions. Git commit history is conducive to accurately and faithfully recording the actions performed with git. However this often can be inflexible and confusing for establishing a clean linear history. That Software Heritage automatically preserves git commits compounds the risk that git commits do not correspond well, or perhaps even prevent, an intended clean linear history. It seems likely that there will be situations where having merge commits and non-linear git commit history will be convenient.

Having edition history separate from git history also provides a potential path for an enhancement akin to retractions of specific editions.

Use of git tree paths instead of git tags

Edition numbers are similar to release versions of software projects. The usual practice in software development is to use git tags to identify releases. In contrast, this specification takes a different approach. Edition numbers are recorded with file paths in git trees rather than git tags.

The main reason for this different approach has to do with how digital successions are recorded and copied. A highly desirable feature for digital successions is the ease of copying without errors. In particular, it is convenient to copy digital successions from multiple sources and store them in a single git repository. In this specification, a complete digital succession is captured by just a chain of commits (initial commit to latest commit). In contrast, entire software projects, which often include release tags,

are usually copied by cloning an entire repository. If edition numbers were recorded as git annotated tags, copying digital successions properly would be a more complicated and error prone (due to missing tags).

Git repository branches are a convenient tool for managing digital successions. But branch names are not part of the digital succession record.

Use of base64url rather than hexadecimal

The textual identifier is in base64url (Base64 with URL and filename safe alphabet) as specified in RFC 4648 [5]. Both this format and hexadecimal have pros and cons.

As mentioned in the introduction to the textual representation, there is no obligation to use the Digital Succession Identifier (DSI) generated by a brand new digital succession. If a DSI is unacceptable to the creator of a new digital succession, a new DSI can be easily generated. Helper tools could perform this function for users by default.

The main con to base64url is that it can contain characters which are more prone to copy errors when transmitted via human sight. Certain fonts and handwriting can make poor or no distinction between certain character pairs. For example, some popular sans serif fonts make no visual distinction between capital I and lowercase l.

The main pro to base64url is that it is 27 characters rather than 40. Given that DSIs are used in contexts similar to DOIs, it is likely more acceptable to adopters to use 27 characters which is very typical for a long DOI, rather than 40 characters. Displaying an ID of 27 characters rather than 40 characters also fits better on web pages rendered on mobile devices. 27 characters are less like to be hidden by showing ellipsis, horizontal scroll regions or shrunk into extremely small fonts.

This design decision is partly made on the belief that the copy-via-human-sight issue is increasingly mitigated by a few technology trends:

1. Computer system are relied upon more and more via more reliable copying mechanisms such as clicking on hyperlinks, copy-and-paste, and QR codes instead of visual copying and handwriting.
2. Websites tend to switch to using appropriate fonts for code like base64url IDs, when output into HTML. The same applies for tools that automatically produce PDFs from computer data.
3. Human-to-computer interfaces, such as input fields of Internet connected software, tend to add features like autocomplete and search disambiguation for typos, when sufficient need arises for such a feature.

Minutiae

- An edition number must be a positive integer strictly less than 65,536 (2 to the 16th power). Edition number zero is reserved for special future use.

Acknowledgments

Thank you to Valentin Lorentz for questions about design choices.

Further Reading

- This specification is heavily influenced by the concept of *intrinsic identifier* and related concepts discussed in [2] [4].
- For a discussion on various concepts and proposed terminology about persistent identifiers, see [6]. In the proposed terminology, a DSI is a persistent identifier (PID) that is "frozen" and "waxing" with both intraversioned and extraversioned PIDs depending on the edition number.

Changes from edition 0.2

- add section about multilevel edition numbering
- add design rationale behind git tree paths instead of git tags
- note future expansion paths for text representation

References

1. (2022) [Git – Wikipedia, the free encyclopedia](#)
2. Cosmo RD, Gruenpeter M, Zacchioli S (2020) Referencing Source Code Artifacts: A Separate Concern in Software Citation. *Computing in Science & Engineering* 22:33–43. <https://doi.org/10.1109/MCSE.2019.2963148>
3. (2022) [Pretty good privacy – Wikipedia, the free encyclopedia](#)
4. Di Cosmo R, Gruenpeter M, Zacchioli S (2018) [Identifiers for Digital Objects: the Case of Software Source Code Preservation](#). In: *iPRES 2018 - 15th International Conference on Digital Preservation*. Boston, United States, pp 1–9
5. Josefsson S (2006) [The Base16, Base32, and Base64 data encodings](#). RFC Editor; Internet Requests for Comments; RFC Editor
6. Kunze J, Calvert S, DeBarry JD, et al (2017) Persistence Statements: Describing Digital Stickiness. *Data Science Journal* 16:39–. <https://doi.org/10.5334/dsj-2017-039>