

## Executive Summary

The Standard Citation Forms Linked Data Working Group (SCF LD WG) was tasked with making a recommendation regarding the scope of the Standard Citation Forms (SCF) resource in the evolving linked data environment. Inherent to accomplishing this, the Group strove to develop an extensible, sustainable and nimble solution that will enable the fulfillment of existing use cases within the current bibliographic encoding environment (i.e. MARC), while simultaneously responding to changing data models and emerging use cases. Essentially, in order to be functional in a linked data environment, SCF resources must have persistent, dereferenceable URIs. The Group identified and evaluated four possible solutions, namely: Reuse id.loc.gov BIBFRAME Works and Instance data; Utilize Wikidata URIs; Reuse OCLC Work entities; or Host locally on rbms.info. The reuse of id.loc.gov metadata and utilizing Wikidata URIs were both deemed viable approaches, while OCLC Work entities and hosting locally are not feasible. The SCF LD WG recommends expanding the scope of the SCF editorial team to encompass the inclusion of persistent, dereferenceable URIs. Currently, it is still necessary to maintain a standardized string for SCF resources.

Our determination is that the reuse of id.loc.gov data is the most viable approach. The stability of the data is likely, as well as having a lighter workload (compared to Wikidata). Additionally, Wikidata is by nature mutable by anyone, which is not desirable for SCF resources. Although some outstanding questions remain regarding implementation, the inclusion of id.loc.gov URIs is a relatively lightweight solution that will enable SCF resources to attain linked data functionality.

## Charge

The SCF LD WG was charged with developing a recommendation to address the following, with a submission deadline of January 31, 2020.

1. Evaluate the function of the SCF resource's citation strings in a linked data environment. This evaluation should include (but not necessarily be limited to) a review of how SCF strings would be expressed in BIBFRAME / the Art and Rare Materials BIBFRAME Extension (ARM) bibliographic citations model. The evaluation may also consider whether SCF citation strings could be used in conjunction with other recording methods for making bibliographic references.
2. Evaluate whether the working principles and scope of the SCF resource should be altered or expanded to best fulfill community needs concerning citations in a linked data environment.
3. Make recommendations for maintenance and development of the SCF resource based on the evaluations described above.

## Approach

As noted in the Standard Citation Forms introduction and in the prefatory material of the group charge, the SCF resource currently provides standardized citations, taking the form of strings, or instructions on how to construct such a string, in order to consistently identify bibliographies commonly used in rare materials cataloging. In order to fully understand and ensure the role of the aforementioned strings were appropriately represented in any solution the group developed, we began by outlining users and use cases for the Standard Citation Forms (outlined in more detail below). The SCF editorial team of the Bibliographic Standards Committee was consulted as the major stakeholders in this effort. Subsequently, the group established minimum functional requirements for a product that would meet both current and future needs. The group then analyzed possible options for realizing that product. Considerations in this analysis included:

- Ability to meet minimum functionality requirements
- Usability of the resulting resource
- Technical feasibility (initial and long-term)
- Sustainability and longevity of the resource and its maintenance

## Outline of users and use cases

### USERS

Standard Citation Forms draws users from across the spectrum of special collections. However, it is useful to consider two primary categories of users for the purposes of this report: end-users of the catalog and users of the SCF resource. Additionally, the needs of the SCF editorial team, as the party responsible for ongoing maintenance, must be considered. End-users of the catalog comprise a variety of demographics, from undergraduates, librarians and scholars. The primary users of the web resource are catalogers, although of varying levels of expertise, as well as booksellers.

### USE CASES FOR THE WEB RESOURCE:

To understand the use cases for catalogers and other users of the web resource, the SCF LD WG turned to the SCF editorial team. Two primary subsets of use cases were highlighted by our feedback from the SCF editorial team: verification and research guidance. Catalogers rely on the standard citation form *string* to reliably direct them to the information in a referenced, usually printed, work that will either justify a piece of information, undeniably identify a work-in-hand, or otherwise provide information about variances in a given resource. In the second case, a cataloger will use the extant citation or the web resource broadly as a gateway to discovery of similar resources. The following list represents quoted use cases provided by the SCF editorial team.

## 1. Verification

- a. When I see a citation already in a catalog record, I can use it to verify that I have the same edition as that represented by the catalog record, if the catalog record does not make that clear, and to verify details about the book in hand, such as signature statements
- b. When I add a citation to a catalog record, I am most often using the citation to verify that the book I have described is the same as the one described in the bibliography I am citing
- c. Another use case when I add a citation to a catalog record is to justify my identification of an anonymous author. I couldn't find a specific example quickly, but if I'm cataloging an anonymous work, and use a bibliography to identify the author and put that author in the 100 field, I then include a 500 note indicating that's where the identification came from, and a 510 with the specific citation
- d. Identify variant bindings
- e. Verify signatures
- f. Distinguish between editions

## 2. Provide topical guidance

- a. A cataloguer or researcher might use the SCF resource to find resources on a specific topic such as incunabula or books printed in Italy, because she's looking at a book that doesn't have a printer listed or is missing a vital page, and she wants to use a bibliography to find and supply that information
- b. A cataloger might use the SCF resource to find the citation form for a bibliography or resource about books on magic or books printed in New Orleans that he wants to reference, whether to demonstrate that he's done his due diligence or to provide his researchers/readers with additional avenues of information about the book in hand
- c. A researcher might use the SCF resource to trace a reference, particularly if that reference were abbreviated or otherwise cryptic, that she's seen given in a bibliography or catalogue. Maybe say, several bibliographies refer to the Wing as having particularly vital information about a title or an author, but she doesn't know what or how Wing might be, and it might take her a while to track down the Short-title catalogue of books printed in England, Scotland, Ireland, Wales, and British America, particularly in some more modern iterations of the library catalogue which can be more bazaar than index

## USE CASES FOR THE END-USER

Use cases for the end-user of the catalog mirror those of the cataloger in some cases, but often represent a broader set of uses. Additionally, because the experience of the end-user of the catalog varies so dramatically based on the initial discovery point and resultant retrieval pathway, it is difficult to gather and articulate uniform use cases. As such, the SCF LD WG

relied on our personal expertise. Importantly, the use cases for the end-user of the catalog are those that have the most potential gain in a linked data environment.

- Confirm and/or verify portion of description
- Be directed to additional information about a described resource contained in a bibliography
- Locate similar or topically related resources based on the citation (i.e. locate other works referenced or described in a given bibliography)
- Learn that an item is not described in a given bibliography

## Minimum functional requirements

1. Users must be able to search and locate resources identified as belonging to the “Standard Citation Forms”
2. Provide straightforward citation string for use in MARC records
  - Because bibliographic models for linked open data have not been implemented in mainstream production, particularly for rare materials, most libraries rely on MARC encoding standards and will continue to require a standardized and easily accessible string to include in MARC records.
3. Compatibility with BIBFRAME and ARM
  - Requires external, actionable URIs at the work level.
  - ARM currently requires URIs for each citation entry, but these do not need to be externally minted.
4. Allow pathway for inclusion of resources that do not have existing OCLC or id.loc.gov metadata.

## Potential Platforms

### Reuse of id.loc.gov data

Id.loc.gov is a product by the Library of Congress (LC) to serve the LC Name Authority File, LC Subject Headings and a variety of other controlled vocabularies as linked data. Further, LC began publishing Work and Instance data on id.loc.gov in 2019, providing as BIBFRAME (BF) the bibliographic resources LC expects libraries to reuse in cooperative cataloging practice. The BF Works and Instances published on id.loc.gov remain somewhat experimental in that this is the first time LC has published their BF data at-scale and the services around this may shift as use cases continue to develop; that said, informal conversations with the BIBFRAME and id.loc.gov developers make the SCF LD WG believe that LC is committed to continuing to provide these data. These BF Works and Instances are created as part of a conversion from MARC in two streams: LC's catalog as well as name/title authorities. Given the extent of LC's holdings and that a name/title authority may exist for others, many of the entries in the SCF resource are likely represented in the extant BF data.

A pathway for adding entries will require additional research. One option would be to create a new name/title authority which will be automatically incorporated into the dataset by LC. This path, however, will not include the creation of an instance record. Instance records are currently required for ARM modelling. Currently, the addition of new MARC records via OCLC numbers provided by SCF to id.loc.gov will not be feasible for LC. Further research is needed to determine the number of affected records in the SCF resource. The SCF editorial group may elect not to include a URI for those records.

It is important to note that the SCF resource would not be a separate vocabulary on id.loc.gov. Since the SCF resource is (in essence) a set of bibliographic data, id.loc.gov developers could not warrant a separate dataset for the SCF entities.

There are advantages and concerns of reusing id.loc.gov data for the purposes of the SCF resource:

- Advantages:
  - The data are in BIBFRAME, the model currently planned for bibliographic description
  - The dataset on id.loc.gov is very extensive for bibliographic data
  - The id.loc.gov developers are very open to continued discussions with the SCF editorial team and interested in understanding how best these data can be useful for the community
  - BSC plans to host the RBMS vocabularies on id.loc.gov
- Concerns:
  - BIBFRAME is an evolving data model; the community does not know how the model may shift as further versions of BF are released
  - Hosting BF Works and Instances on id.loc.gov is relatively recent; while the developers anticipate that this will be stable, the track-record for these particular entities is not long-running. That said, there is no reason to question LC's commitment to this platform or these data.
  - It is unclear how best to represent that a resource (BF Work or Instance) is part of the SCF resource. The SCF LD WG members discussed the possibility of adding data to the SCF-related resources denoting their inclusion in the SCF database; however, that conversation did not yield anything conclusive and further conversation is necessary.
  - A straightforward pathway for including non LC resources will need to be developed

Id.loc.gov represents a good source of reusable data. Additional user experience requirements for navigating the SCF resource would need to be facilitated on rbms.info, including the creation/continuation of the Author, Title and Subject lists, if desired. Two paths have been identified during the course of the group's work: flat lists on the website with links to id.loc.gov OR a database as currently exists with the addition of URIs linking to resources on id.loc.gov. If flat lists are selected, the maintenance of the SCF resource would be greatly reduced; however,

the browse functionality of the current website will be lost. Alternately, the SCF editorial team, in conjunction with the web team, could elect to continue to maintain the existing database functionality and incorporate id.loc.gov URIs into a newly added field. URIs could likely be matched automatically with SCF resources to reduce manual data entry.

## Wikidata

Wikidata is a centralized place where structured data is stored for wikis such as Wikipedia and Wiktionary. It is freely licensed and can be used as linked data for a variety of projects. Every item or concept in Wikidata has an identifier (the number prefixed with 'q') and an actionable URI can be obtained by attaching that ID to the Wikidata concept namespace. Wikidata editors can create entries for anything, including works, manifestations, even down to specific items.

The Standard Citation Form editorial team could make use of Wikidata by adding the URI to the entries in the list on the website or could potentially use the data to populate into a different database, similar to the two paths identified for use of id.loc.gov previously mentioned. Regardless, because the wiki storage repository is so big, the data specific to the SCF resource would still need to be listed in some other contained form outside of Wikidata alone. The tasks of the SCF editorial group would be to add relevant data to Wikidata. Currently, there isn't a strong book presence in Wikidata so this would involve adding a lot of data. Edit-a-thons are a popular way to get help & improve community engagement, but all the entries we currently have in the SCF resource would need to be added. URIs could then be added to the list on the SCF website.

For maintenance, the SCF editorial group would need to add new entries to Wikidata. There is some concern that because Wikidata can be edited by anyone with an account, validity of data is not secure. Should they choose this route, the SCF editorial team might want to periodically check for accuracy of data as well. This means that as far as technical requirements go, the SCF editorial group would need to learn how to input data into Wikidata and, as mentioned before, quite a bit of data would need to be added at the onset.

Because Wikidata serves as the basis for much of Wikipedia's authority data, the assumption is that Wikidata is going to remain a constant resource into the foreseeable future. There is, however, the chance that relevant data gets changed by other Wikidata community members. Wikidata is a viable option; however, because of the set up involved and the fact that the library community does not have full control over the data, we do not recommend it as the best option at this time.

## OCLC Work Entities

Part of the WorldCat Entities program, WorldCat Works provides actionable, persistent URIs for “work” entities, which are analogous to the FRBR concept of a work. The program is focused on providing URIs for work records and does not provide URIs for any other level (e.g. expression, manifestation, item).

URIs created as part of the program live in the “experimental WorldCat Linked Data” located in WorldCat bibliographic records and are accessible via the “exampleofWork” property. The page documenting the WorldCat Works program states that URIs are also accessible via OCLC’s xID web service. However, OCLC [decommissioned xID services in 2018](#).

OCLC introduced WorldCat Works in 2014 and indicated that it would continue to develop the program, improve linked data access and create work record URIs. However, when the SCF LD WG began examining WorldCat Works as an option in 2019, few URIs in the experimental Linked Data section of the record were functional. Attempts to access work records frequently resulted in a standard “not found” error message (HTTP 404). This error message appears for bibliographies commonly used by catalogers like [Adams](#) and [Blanck](#) as well as for popular, widely read works. The URIs found in this program are thus not actionable. The SCF LD WG reached out to OCLC with questions regarding the program but did not receive a response.

It is unknown what the user experience associated with WorldCat Works would look like. If the URIs were functional, the SCF editorial team would have three options: 1) Add the work URI to an existing entry on the SCF site, thus replicating the current SCF experience; 2) Re-build the current site for the dynamic population of data from the OCLC URI (specifically author, name, description, subjects, etc.), and as a work entity record was updated, the SCF resource would then automatically update; 3) Use OCLC as the sole host for SCF resources. This third option would be detrimental to user experience as locating both the appropriate bibliographic record and the URI in the experimental section of that record is fairly time consuming and might lead to less-than-ideal searching scenarios.

Setup and ongoing maintenance associated would be dependent on which hosting option the SCF editorial team selected. If the team decided to add URIs to existing entries, volunteers would need to search for and add those links to entries; it may be possible to do this work in batch, but that is unclear at this point in time. Dynamic population on the existing SCF site would require extensive work by the RBMS Web Team and possible collaboration with OCLC. It is not known what collaboration with OCLC would look like.

Since URIs in OCLC are not actionable, the longevity of the data cannot be ensured. The LD WG does not recommend using the WorldCat Works program.

As a note, in early 2020, OCLC announced that they received a [Mellon grant to support linked data management initiatives](#) involving work and author entities. The news of this grant came out after the LD WG concluded the substantive body of their work. As such, there was no further examination of OCLC as an option.

## Internal hosting via rbms.info

The SCF LD WG also considered internal hosting via rbms.info but ultimately found this to be infeasible for a variety of reasons. Due to the inflexibility of the wordpress platform, rbms.info is incapable of minting actionable URIs, thus taking it out of the running as an option that can fulfill the minimum requirements. The user experience of the SCF resource for catalogers is impossible to assess in this scenario given its aforementioned infeasibility.

Theoretically, SCF editorial tasks would mirror those outlined above, but with the addition of technical infrastructure maintenance that surpasses both existing and projected capacity and expertise. Because of the impracticability of this approach, the questions of what necessary resources and the data's longevity are moot.

## Recommendation

Of the four options investigated by the SCF LD WG, only two meet the minimum functional requirements for representing bibliographic citations: Wikidata and id.loc.gov data. With both options, the SCF editorial team can add actionable, externally minted, persistent URIs to bibliographic resource entries in the current SCF resource without significantly increasing the responsibilities of the SCF editorial team and RBMS web editors. However, both options involve ceding some amount of control over the data represented in SCF to outside stakeholders, namely Wikidata editors and Library of Congress. In the case of Wikidata, ceding control means that any data input by the SCF editorial team into Wikidata is inherently mutable and open to editing by any interested editor. Data in id.loc.gov is maintained by the Library of Congress, a known entity and a partner already involved in BSC linked data initiatives.

The SCF LDWG thus makes the following recommendations related to the representation and maintenance of bibliographic citations as linked data in the SCF resource:

1. We recommend including persistent URIs linking to id.loc.gov entities on all resource entries in the SCF resource to fulfill the requirements set out by the ARM bibliographic citations model. By using data made available via Library of Congress, the stakeholders are all known entities already deeply involved in the creation of library related metadata. Taking this approach allows the SCF editorial team to make use of readily (and already) available data via BIBFRAME Works and Instances created by LC via MARC bibliographic metadata conversion from their catalog as well as the conversion of the name/title authority file. For SCF resources not held at the Library of Congress,



further research will be required to develop a workable solution. More information on the possible solutions/approaches can be found in the id.loc.gov section above.

2. We therefore recommend that the working principles and scope of the SCF resource be expanded to include adding those URIs to existing entries in the SCF database.
3. We recommend reconvening the SCF LD WG as major developments related to linked data in library communities arise. Recommendations #1-2 reflect the iterative nature of linked data development and what is currently possible given technological constraints. Including external URIs on entries in the SCF resource allows the SCF editorial team to continue fulfilling the existing needs of its users and use-cases, while adapting to likely future needs. As linked data continues to evolve in library communities, there is an ongoing need to evaluate practices and discuss future directions, potentially including (but not limited to) the long term maintenance and use of the existing SCF resource and the data hosted in id.loc.gov.