Share-VDE (SVDE) PCC Data Pool Evaluation Task Group

Report, March 3, 2023

Background and Scope

During 2021, and in the context of the Linked Data for Production phase 3 (LD4P3) grant, the Share-VDE initiative worked with PCC and OCLC to bring all BIBCO and Conser MARC records into an open pool of PCC-quality BIBFRAME data, hosted in a dedicated data store of the Share-VDE infrastructure. The PCC data pool is a beta version that requires testing and review to expose data conversion issues and is accessible at <https://pcc-lod.org/>.

The Share-VDE (SVDE) PCC Data Pool Evaluation Task Group was formally charged by PoCo on October 15, 2022, to 1) generate a plan to review the accuracy and completeness of PCC bibliographic data available in the Share-VDE PCC catalog Beta

version, 2) gather input from the PCC community and coordinate review activities, and 3) consolidate community feedback for delivery to Casalini. The group was not asked to provide feedback on authority data, user interface, navigation, etc.

All documents, forms, meeting notes, and recorded sessions were organized and saved in the [Google Drive Folder](#) that was accessible by TG members.

In the following report, the Task Group (TG) discusses the charge and deliverables and how they have been achieved.

Develop a plan to coordinate PCC data review

PCC Testers Pool

The TG decided that the best way to involve the PCC Community in the review of the PCC data pool was to have an open call for reviewers. Aiming to ensure wide representation, the TG developed a volunteer sign-up form asking for information regarding cataloging domain and language/script expertise. The call for volunteers was launched on November 2, 2022, and it invited different cataloging communities to join the testing using the following distribution lists: PCC, BIBCO, CONSER, BIBFRAME, LD4, and LD4 non-Latin Affinity Groups lists. The call was also shared on the LD4 community Slack workspace. Given the short timeframe for the review process, the deadline to express interest in participation was November 11, 2022.

Volunteers Composition

The call for volunteers had 53 responses with different levels of commitment for the review: "at least 4 titles" (~54%); "5-10 titles" (~ 29 %), and "10 or more titles" (~16%). Participants were from different types of institutions, mostly academic, but vendors, national, and public libraries were also represented. Respondents were mostly from the United States, though there was also representation from institutions outside of the United States, including University of Alberta (Canada); University of Lagos (Nigeria); National Library of Nigeria Kwara; National Taiwan University Library; Fundación Ignacio Larramendi (Spain); and National Library of Spain.

The volunteers expressed interest in covering the following type/format of materials: general print (~40%); audiovisual (~23%); music (~10%); rare books (~24%); serials (~35%); special collections (~5%); non-Latin languages (~32%).

The language/scripts represented included: English, German, Dutch, French, Spanish, Hungarian, CKJ, Arabic and Persian, Hebrew/Yiddish, Serbian, Scandinavian, all Slavic

(Russian, Belarusian, Ukrainian, Serbian, Croatian, Polish, Czech, Slovak, Slovene, Bulgarian, Montenegrin, Bosnian), and non-Slavic in Cyrillic script (Central Asian languages: Kazakh, Tajik, Uzbek, Kyrgyz, Turkmen).

Training and Documentation

Because not all members were familiar with the Share-VDE platform, the TG set up a walkthrough of the user interface led by the Share-VDE Team on November 16 as a kick-off of the review period. Around 40 testers attended this session. Anna Lionetti, R&D Assistant at Casalini Libri, provided an overview of the Share-VDE initiative and the PCC data pool. Filip Jakobsen, user experience lead for Share-VDE, gave a live walkthrough of the site followed by a live Q&A session. The session was recorded for members who were not able to attend and for future consultation. To complement the walkthrough session, the TG also created training documentation with instructions, hosted on a dedicated Google drive folder shared with the volunteers (See Appendix 1).

Additionally, and to coordinate communication with and among testers, the TG set up the Google group pcc_svde_catalog@googlegroups.com.

Gather input from the PCC community and coordinate review activities and evaluation tools

To collect the testers' feedback about the bibliographic data in the PCC data pool, the TG developed an evaluation Google form (see Appendix 2) and provided instructions documenting the scope of the review process and the use of the form. The TG discussed the different options to adequately represent the reviewer's feedback, and eventually decided to have a brief and easy-to-fill form.

Volunteers were asked to review at least 4 bibliographic descriptions in the PCC Catalog for both the original work (bf:work) and the publication (bf:instance), and were prompted to compare the Share-VDE BIBFRAME description with the data available in the OCLC PCC record, focusing on the accuracy and completeness of the data. Authorities and usability or navigation of the platform were out of the scope of the review. Volunteers were expected to fill out one evaluation form for each bibliographic description they reviewed, regardless of their findings (even if no issues are identified). This has helped the TG to get a better picture of what has been reviewed in total, and put the issues identified in a larger context.

To reinforce the training and make sure that all testers were doing well with the review process, the TG scheduled two office hour sessions to assist testers in conducting the review efficiently and to address any issues they might encounter. The office hours were scheduled on December 2 and December 15, and they were recorded to benefit members who were not able to attend and for future consultation.

The review period ended on December 31, 2022.

Consolidate community feedback for delivery to Casalini

Data Analysis

The TG received 212 evaluation form submissions covering different aspects of the PCC data pool: display of bibliographic data, relationships between works and publications, and relationships between agents and works or publications. The data tested was related to different formats: monographs, serials, rare book materials, electronic resources, audio visual materials, music, and maps. Materials in different languages and scripts were also represented in the data: English, French, Spanish, Danish, German, Chinese, Japanese, Latin, Greek, Arabic, Hebrew, Russian, Cyrillic. Out of the 212 submissions, 17 did not report any issue.

To organize and consolidate the feedback in a meaningful manner, the TG parsed each form submission into discrete issues and categorized the issues based on an ad-hoc category list developed by the TG during the analysis process. This resulted in an inventory of 427 categorized issues. The following list provides some information about the categories identified, with the number of related issues in parenthesis.

- **Missing MARC fields** (89): Comments related to the absence of data provided in a MARC field. Notice that there are other categories that address specific MARC fields, which might be missing, such as subjects and notes
- **Description issues** (78): A very general category containing all description issues that don't fit in any of the more specific categories
- **Notes fields issues** (71): Because there was a considerable amount of comments about the notes field, the TG decided to group them under a separate category
- **Linking to agents issues** (37): Comments addressing the relationships between agents (persons, organizations) with works or publications. Examples of this include a contributor linked to the wrong work, as well as missing relevant agents

- **Display of vernacular scripts** (43): Comments addressing the presence or display of metadata in non-Latin scripts
- **Missing subjects/genres** (33): Comments about missing LCSH, genres, and classification numbers
- **Linking publication/work issues** (27): Comments addressing the display/presence of relationships between original work and publication, as well as related works
- **Usability issues** (23): Although usability was out of scope, it is hard to separate usability issues from data issues. This category considered all comments related to findability and navigation, as well as labeling issues related to specific formats
- **404 - Page not found** (13): Reviewers have found a number of broken links between entities leading to a "404 - Page not found" error
- **Display of agents** (11): Comments regarding the way the agents are displayed in the data pool
- **Entity duplication issue** (5): Comments regarding duplication of work or agent entities
- **Translations** (4): Comments addressing the description and linking between original work and translated work

Communication with Casalini Libri and the Share-VDE team

The TG worked with Anna Lionetti to deliver the community feedback to Casalini Libri and the Share-VDE team. For this purpose, the TG moved all the data to a Google sheet provided by Share-VDE, including not just the reviewers comments but also the issue category, the language/script of the resource and its format when evident from the comments. The hope is that this additional data can help the Share-VDE team in the review of the feedback.

The tester's feedback was finally delivered to Anna Lionetti on February 2, 2023. The feedback document shared with Share-VDE can be accessed here: <https://docs.google.com/spreadsheets/d/1GFGLsIRFFys749qv7I4G_iZ53Ip8z91WsfQx3Fcn7Ww/edit#gid=1434435994> (Tab: PCC testers feedback).

Communications between Share-VDE and the TG still continues as Share-VDE reviews the feedback provided. The TG will work on addressing any question coming from Share-VDE regarding the feedback and, if needed, will put Anna Lionetti in contact with specific reviewers for further information.

Feedback regarding the review experience

On February 8, 2023, the TG hosted one last office hour to share thoughts about the evaluation experience.

The office hour was attended by 14 of the volunteers. They were asked about their initial proficiency level with the Share-VDE catalog, BIBFRAME and linked data, as well as their experience reviewing the PCC data pool and the support received from the TG.

Of the 14 office hours attendees, almost half of them identified themselves as absolute beginners with the Share-VDE catalog, while the level of expertise with BIBFRAME and linked data concepts was more distributed between beginner-medium, medium and advanced with just very few absolute beginners. While the majority of attendees reported a good evaluation experience, a few regretted not being able to look directly at the underlying RDF data to get a better insight into Share-VDE BIBFRAME, which was outside the scope of this review. Expectations management was also reported as an issue, since volunteers were not sure of what data was not currently represented in the data pool (e.g. notes and subjects) versus what might have been an issue related to data transformation or a lack of understanding of the data model by the reviewer.

The feedback form was also discussed as sometimes too flexible and sometimes not flexible enough to represent some of the issues encountered. The TG understands this is due to the wide range of potential types of issues, and believes that could be avoided in the future with more targeted reviews/forms. Further training on how to conduct the review, as well as a walkthrough of the Share-VDE entities and how they relate to each other, were also identified as potential improvements for future reviews of these characteristics. On the other hand, those present in the meeting agreed that the office hours were extremely helpful to ask questions and share thoughts with the TG and the rest of volunteers, and appreciated the fact that the responses spreadsheet generated by the form was available for their review so that they could have a bigger picture of what types of issues were being reported by the group.

Finally, meeting attendees shared the positive outcomes of their experience reviewing the PCC data pool. It was brought up that looking at bibliographic data converted to BIBFRAME and in a linked data environment sparks thoughts and questions about our current cataloging practices in MARC (e.g. complex subject heading strings) as well as what additional data we could start adding to our bibliographic descriptions to fully leverage its future navigation and discovery in a linked data environment.

To complement the feedback gathered during the office hour and to allow more members to share their reflections on the testing experience, the TG distributed a follow

up survey including the same questions discussed during the last office hour. The 17 responses (see Appendix 1 & 3) that were received confirm the above mentioned feedback provided during the office hour (see appendix for metrics). From the qualitative questions we can gather that those testers which did not think that the Share-VDE walkthrough was sufficient to prepare them for the testing, would have liked more demonstrations, and information regarding the SVDE platform limitations, such as the absence of subject data. For improving the Feedback form, comments from testers varied. While some testers were satisfied with using the form to document their observations, others suggested having more instructions and examples and assigned searches.

Overall, testers who either participated on the last office hour or filled out the follow up survey were appreciative of the opportunity to test the PCC data pool and practice searching the data in a linked data system. However, some challenges were noted, some related to the system itself and its functionality; such as lack of access to 'raw' RDF entity descriptions and lack of a published ontology. Others were related to timing of the testing.

Conclusions and next steps

There are a number of factors that have been challenging in the coordination of the PCC data pool evaluation and feedback consolidation. The TG recognizes that they would have liked more time to develop a more robust feedback form and evaluation workflow but due to the time constraints and the desire to get the volunteers involved as soon as possible, it was decided to opt for a more open ended feedback form and flexible workflow that would then be adjusted based on discussions with the reviewers during the office hours. On the other hand, because the TG was not fully familiar with the Share-VDE BIBFRAME implementation and current status of the PCC data pool, it was challenging to manage reviewers expectations in terms of what constitutes an issue. The decision was to ask reviewers to report anything that was not accurate or missing, in the interest of allowing the Share-VDE team to determine whether the feedback was actionable or not. The TG would have highly benefited from discussing with the Share-VDE team the status of the data in the PCC data pool, before getting the reviewers involved. For instance, knowing in advance that subject access or descriptive notes were not yet fully implemented would have helped the TG plan a better workflow approach.

Volunteers have also expressed that they would have liked more time to get familiar with the platform and further guidance during the review process. In order to make the

review as straightforward as possible, volunteers were prompted to compare an OCLC PCC MARC record with its representation in Share-VDE and report any accuracy or completeness issue, but as the process unfolded, there were questions about whether this was an efficient way of reviewing the data transformation.

On the other hand, many volunteers expressed having a positive experience related to exploring a discovery platform for bibliographic linked data and enhancing their understanding of BIBFRAME and linked data. This leads the TG to believe that tools like the Share-VDE PCC data pool could be leveraged for training purposes by the PCC community.

Task Group Membership

- Iman Dagher (UCLA), co-chair from SCS
- Paloma Graciani Picardo (University of Texas at Austin), co-chair from LDAC
- Nancy Fallgren (National Library of Medicine), PCC SVDE liaison
- Christine Fernsebner Eslao (Harvard)
- Jim Hahn (University of Pennsylvania)

Appendices

Appendix 1. Links to relevant evaluation documentation

- Evaluation Form - https://docs.google.com/forms/d/e/1FAIpQLScQup1NtrHGKz8F2bYvWQYlxehoovwLQ2no-55YrRdeWcm7kQ/viewform
- Share-VDE PCC Data Pool Evaluation Instructions - https://docs.google.com/document/d/1hOtcA2JVTdTBOQCoevXm5cJcfXDjXZMD-36X4I1_M4U/edit#heading=h.szj6jrh8jtwa
- Share-VDE PCC testing Q&A - https://docs.google.com/document/d/1cPczpV2j8JvTp_itD80LZlCi1L8Y_rPtcqHgk-WUH8g/edit
- Follow up Survey Form https://docs.google.com/forms/d/1N7MrfzzspF6qixTiGbGurblSNTuMVthRNSnIJf85D8Q/edit

Appendix 2. Evaluation form

The following is the text of the evaluation form:

> The goal of the Share-VDE PCC Data Pool Evaluation is to assure that the PCC bibliographic descriptions are adequately represented in the Share-VDE [PCC Catalog](#).
>
> Fill out this form for each of the bibliographic descriptions you are reviewing (**work** and **publication** level).
>
> You can find more information about the evaluation process and how to fill out this form on the [Share-VDE PCC Data Pool Evaluation Instructions](#).
>
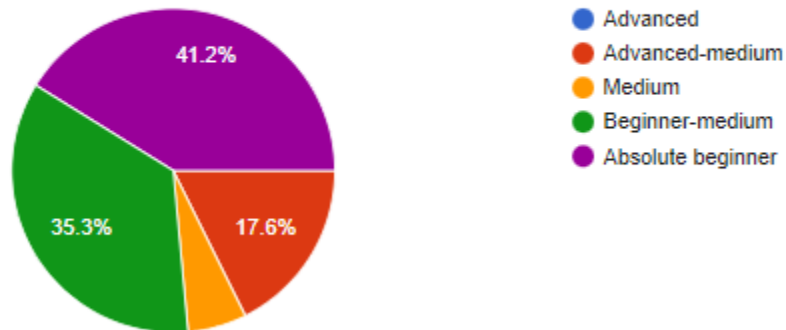> Feel free to post any questions that you might have regarding the evaluation process to [pcc_svde_catalog@googlegroups.com](#).
>
> *Thanks for providing your feedback!*
>
> - **Your name** (This will help the task group follow up with you if needed
>
> - **Language/script** (Indicate language/script of the resource you are reviewing or looking to review
>
> - **Entity URI** (Check the review instructions for information on where to get the URI: [Share-VDE PCC Data Pool Evaluation Instructions](#)
>
> - **Have you found any issues with this description?** (Yes/no)
>
> - If you have found any issues, please describe them below, otherwise, you can submit the form now. Thanks!
>
> - **Description of the issue** (Briefly describe the issue in a couple sentences. Please include (1) entity type (work / publication / agent); (2) nature of the issue (completeness / accuracy / other); (3) expected results of the review; (4) actual results of the review.)
>
> - **Notes** (Please add any other information that would help define the review issue)
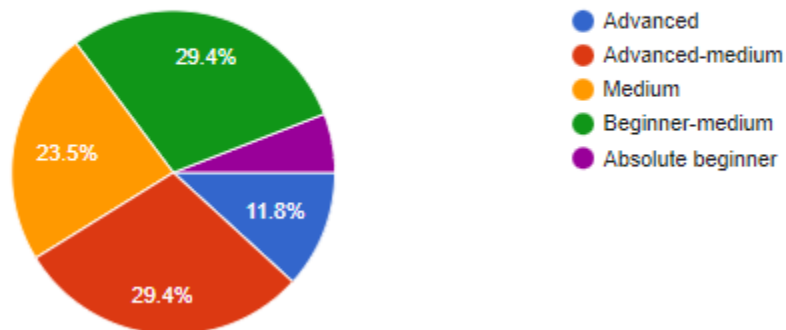
Appendix 3. Metrics from the follow up survey results

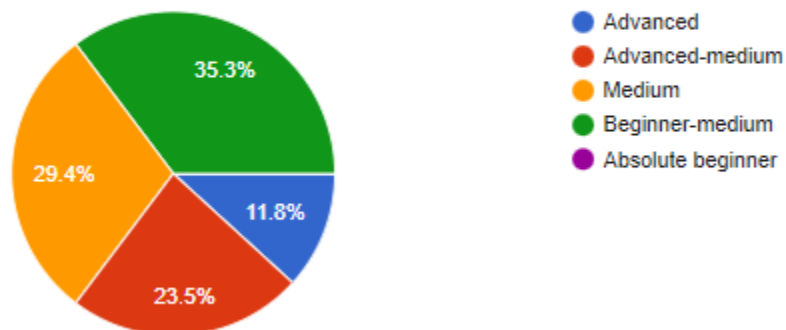**Familiarity level with Share-VDE Platform before the testing period**

17 responses

- ● Advanced
- ● Advanced-medium
- ● Medium
- ● Beginner-medium
- ● Absolute beginner

2. Familiarity level with BIBFRAME

17 responses

- ● Advanced
- ● Advanced-medium
- ● Medium
- ● Beginner-medium
- ● Absolute beginner

3. Familiarity level with Linked Data

17 responses

- ● Advanced
- ● Advanced-medium
- ● Medium
- ● Beginner-medium
- ● Absolute beginner

4. How was the support provided by the Task Group for the PCC data pool review?

17 responses

- ● Very good
- ● Good
- ● Indifferent
- ● Bad
- ● Very bad

47.1%

41.2%
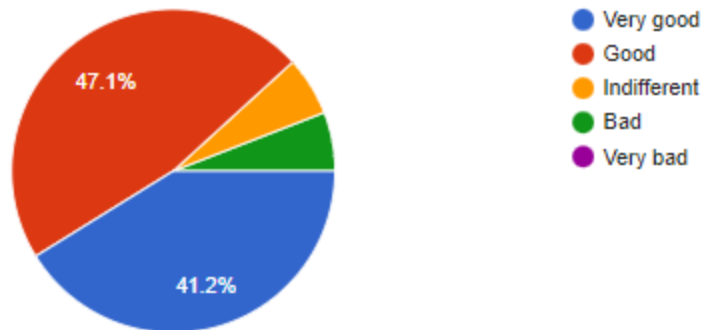
5. Was the walkthrough provided by Share-VDE sufficient to start testing?

17 responses

- ● Yes
- ● No

11.8%

88.2%
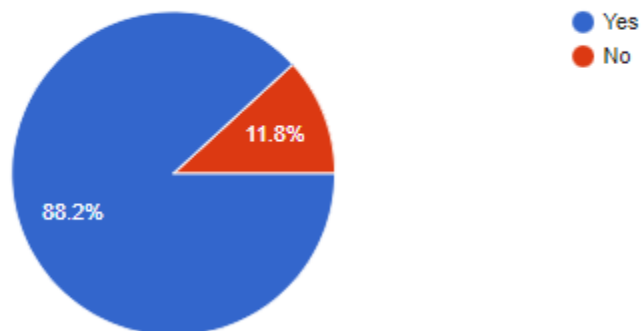
6. Were you able to add your comments efficiently in the form?

16 responses

- ● Yes
- ● NO

12.5%

87.5%