PCC and the transition to Linked Data - implementation during LD4P3

PCC Joint Operations Committee Meeting, May 7th 2021

Michele Casalini
michele@casalini.it

https://wiki.svde.org
PCC data pool progress update

- **Initial load** from OCLC to SVDE by 31st December 2020 → approx. 4.5 million WorldCat MARC bibliographic records created by PCC libraries (042 ##$apcc)
- **SVDE** has delivered PCC records converted in BIBFRAME along with the original MARC records enriched with URIs
- Converted records delivered to OCLC through ad hoc pipeline; also available here [PCC](#) along with MARC records enriched
- **Regular updates** are ongoing, on monthly basis, for the duration of LD4P3 to follow, PCC entities will be available in the CKB
- **PCC data** on SVDE portal and/or on PCC dedicated skin in preparation
Conversion, enrichment, reconciliation and housing of PCC data into the Share initiative as autonomous tenant with a separated enriched CKB and local PCC URIs.
Inclusion of PCC data into Share-VDE as tenant: first step

https://svde.org/pcc/rdfBibframe/Agent/1234

SameAs
http://share-vde.org/sharevde/rdfBibframe/Agent/38006

SameAs
http://isni.org/isni/0000000121032683
Workflow PCC - SVDE

1. The PCC can be considered as a new tenant to be included in the Share Family;

2. the PCC MARC data feed the clusterization tool Authify;

3. enrichment of the PCC MARC data through Authify;
   - URI enrichment from various data sources: LC, GND, VIAF, OCLC, Wikidata, Share-VDE;
   - SVDE is an external data source serving URI enrichment for PCC;

4. result:
   - “ad hoc” PCC Cluster Knowledge Base including SVDE, LC, GND, VIAF, OCLC, Wikidata URIs (and URIs from other sources);
   - the PCC MARC data are data enriched;
Workflow PCC - SVDE

5. the PCC MARC data enriched feeds
   ○ the Cluster Knowledge Base;
   ○ SVDE platform (through SOLR);
   ○ Lodify conversion tool (conversion from MARC to linked data);

6. PCC data has a specific URI to identify the entities of PCC graph, e.g.
   https://svde.org/pcc/rdfBibframe/Agent/1234
   SameAs
   http://share-vde.org/sharevde/rdfBibframe/Agent/38006
   SameAs
   http://isni.org/isni/0000000121032683

7. PCC data is enriched with Share-VDE URIs (and other sources);

8. Preconditions are set for PCC’s data publication on the SVDE entity based discovery interface.
Benefits of having PCC data in SVDE

**Transition to linked open data:**
- PCC data converted in RDF (both new data and old records)
- original MARC records enriched with URIs

**Persistent identification and authoritiveness:**
- enrichment of PCC data with IDs from authoritative sources (SVDE, ISNI, VIAF etc.)
- ad hoc namespace for PCC URIs → this makes the PCC itself an authoritative source
- enhanced outreach of the PCC and prominent role in the international community

**Exchange of information:**
- easier integration of PCC data in external environments through the PCC URIs
- PCC data are treated according to the SVDE entity model that enables interoperability
Facilitate interoperability between entity models

SVDE Advisory Council decision of June 10th 2020:
Resource that is a svde:Opus is also a bf:Work  --  Likewise resource that is a svde:Work is also a bf:Work
See the SVDE entity model compared to BF and LRM and an example of application of the model
Approach to tenant infrastructure

- Each tenant has its own CKB (e.g. Share-VDE CKB, PCC CKB, Kubikat-LOD CKB etc.)
- Each entity has its own URI in the different CKB namespaces
  - E.g. Antonio Vivaldi URIs in different CKB namespaces [the following URIs are for simulation purposes]
    http://share-vde/agent/123456 sameAs
    https://svde.org/pcc/agent/7890123 sameAs
    http://kubikat/agent/456789 sameAs
- Central index able to point to all the URIs in all the CKBs of the different tenants: SFI - Share Family Index
  - sameAs relationships between URIs of entities in the various CKBs
  - the SFI ID (Share Family Index ID) can be the unique identifier aggregating URIs specific to each CKB, carrying the minimum amount of data needed to identify the object
- Having a “central ID” like the SFI ID that aggregates URIs for the same resource from different CKBs could facilitate a range of additional services across the projects of the various tenants
- Benefit of the SFI: maintain identity of individual project, but also cooperate and exchange with others
Participation and autonomy in the Share Family

Share-Music  Share-Catalogue
Share-VDE  Share-Art
PCC data pool  Kubikat-LOD
Ancient books  Parsifal

Share Family Index
Community engagement: World Wide Web

Mixed community: cross-domain cooperation across the Web community

Scientific value: same solutions serve scopes of different communities, data reuse
The evolution over time: towards SVDE 2.0

Optimisation of LD workflows in production
Conversion from MARC to RDF using BIBFRAME and other ontologies
SVDE authoritative Cluster Knowledge Base
Original MARC records enriched with URIs from different sources
RDF data enriched with URIs from different sources

Infrastructure, data storage, indexing and queries
Internal Postgres RDBMS
Triplestore with SPARQL query endpoint
Solr inverted index
Tenant infrastructure with Share Family Index across Share Family CKBs

Data publication and exchange
Union catalogue and advanced entity discovery platform
API layer for CRUD operations
Interoperability layer with external systems (e.g. LD4P)

Advanced services
New generation authority control
Entity editing tool for update and management of linked data entities: J.Cricket editor
Cross-tenant services
Apply Share principles to other domains (e.g. archives, museums, music)

More on the latest achievements
SVDE Sapientia CKB ecosystem
Authority services: automatic processes

For record environments:
→ MARC record validator
→ MARC corrections for errors and obsolete forms
→ MARC matching/enrichment with SVDE and external profiled sources
→ reporting of MARC records elaboration
→ creation/update and delivery of Authority records to the library

For RDF environments:
→ Access point enrichment (including Series and Subjects)
→ Matching, import and interaction with the Sapientia Cluster Knowledge Base
  (Enabled through the LOD Platform)
Authority services: manual processes

→ Manual control for similar matches and for non matches
→ Entity Work and Agents (including Publisher) access point management
→ Integration with the ISNI registration processes

(Enabled through the URI Registration Platform)

→ Cluster Knowledge Base entity management for Works and Agents

(Enabled through the J. Cricket CKB Editor)
LD4P3 - SVDE closing the loop

- first step done → **API pipeline** that pulls records from Sinopia to SVDE
  - implemented by Sinopia team and tested by SVDE
- now working on **interoperability of the data**
  - important: see how closely the PCC Data Pool RDF matches the RDF in Sinopia
  - SVDE entity model compatibility: svde:Opus and svde:Work are both a type of bf:Work
- exchanges ongoing: **technical meetings are in course**
  - demo of the current SVDE back-end search API has been done, and specs shared to give Sinopia/QA updates about what data will be available for queries and how
  - fine tune data exchange: the evolution of SVDE CKB 2.0 data structure has been shared, work in progress to define steps needed to close the loop
Thank you

https://wiki.svde.org