How to integrate library data into GoTriple: a collaboration with the Text+ Consortium[1]

Luca De Santis
ORCID: https://orcid.org/0000-0003-0527-840X
Net7 Srl, Italy
desantis@netseven.it
Eva-Maria Gerstner
ORCID: https://orcid.org/0000-0001-9920-4238
Max Weber Foundation, Germany
gerstner@maxweberstiftung.de
Reinhold Heuvelmann
German National Library, Germany
Sy Holsinger
ORCID: https://orcid.org/0009-0005-8978-225X
OPERAS Research Infrastructure, Belgium
sy.holsinger@operas-eu.org
Nanette Rissler-Pipka
ORCID: https://orcid.org/0000-0002-0719-9003
Max Weber Foundation, Germany
rissler-pipka@maxweberstiftung.de
Angela Vorndran
ORCID: https://orcid.org/0000-0001-7162-9875
German National Library, Germany
A.Vorndran@dnb.de

Libellarium 16 (1) 2025: 39–48
Professional Paper / Stručni rad
UDK: 001.103.2:025.321.4
Received / Primljeno: 30. 08. 2024.
Accepted / Prihvaćeno: 03. 11. 2025.
doi: https://doi.org/10.15291/libellarium.4541

Abstract

Purpose. This report presents a showcase for the collaboration between a national and a European research data infrastructure and community service. First, the organisational

[1] The paper was originally presented at the OPERAS conference that was held in Zadar in April 2024.

and technical approaches are explained. In the second part, examples for data integration are proposed together with participants. Finally, possible transfer and application of the presented workflow are discussed.

Approach/Methodology. Following a joint analysis by the Text+ and GoTriple teams, developing a dedicated MARC 21 connector for SCRE was identified as the optimal strategy for integrating Text+. This enabled the import of SSH publications from the German National Library (DNB) into GoTriple. This decision also potentially facilitates the integration of relevant SSH collections managed by other libraries, as MARC 21 is an important standard for library cataloguing and data exchange.

Results. There have been several technical challenges in this integration. First of all, MARC 21 is a complex protocol, not easily processed. In this case, the use of dedicated, open-source software libraries was particularly helpful. In addition, multiple options had to be evaluated and tested before a suitable solution for harvesting DNB in GoTriple, as presented here, was found.

Originality/Value. The presented example of GoTriple and one of the Text+ resource providers (German National Library) can serve as a best practice for integrating more library catalogue data into the GoTriple platform. It also provides a model for the essential communication and collaboration workflow needed to estimate the effort required for future data integration into GoTriple at the level of an exemplary national infrastructure.

KEYWORDS: bibliographic data, data discovery, data integration, GoTriple, library catalogue, metadata, OPERAS research infrastructure, social sciences and humanities, Text+

1. Preliminary thoughts on European and national research infrastructures

Linking national and European infrastructure components is one of the most challenging tasks when building research data infrastructure at both levels. Technical interoperability which is mostly given using standardised protocols and schemas (see below and De Santis 2023) is not sufficient on its own. Connecting both levels through an organised knowledge exchange is necessary, especially in the rapidly expanding research infrastructure landscape in Europe.

Apart from the technical challenges and efforts required to set up the necessary platforms, registries and repositories, infrastructure is built by people working in different projects and contexts. It is shaped by various funding programmes, disciplines, or regions, which may result not only in complementary structures but also in duplication. The latter is sometimes necessary to make research data and outcomes findable and visible in different contexts (Dumouchel 2022; Larkin 2013). Language, social and legal barriers prevent users from accessing resources, such as those of national libraries. Collaboration in another context requires preparatory and communicative work not only from the users' perspective but also for infrastructure providers. Therefore, to make finding and using research outcomes and resources easier for the community, providers need to invest in collaboration such as

that presented in this report.
An example of the European level infrastructure that promotes collaboration and combination of various resources is the GoTriple platform. GoTriple is part of the OPERAS infrastructure and as a "young" and developing discovery service, it is continuously improving along with the integrated resources. The initial step for the GoTriple team to provide technical and content-related documentation (De Santis 2023) for resource providers. The next step is the current collaboration and exchange between the service provider and resource provider.
Text+ is a consortium of the German National Research Data Infrastructure (NFDI) dedicated to text and language-based research and is one of the four humanities consortia that signed a memorandum of understanding to collaborate (Brünger-Weilandt et al. 2019).
The presented example of GoTriple and one of the Text+ resource providers (the German National Library) can serve as a best practice for the integrating more library catalogue data (in MARC 21 format) into the GoTriple platform. It also provides a model for the essential communication and collaboration workflow needed to estimate the effort required for future data integration into GoTriple at the level of a national infrastructure.

2. The German National Library (DNB) data catalogue and data centre in Text+

The German National Library[2] is Germany's central archival library. It is mandated by law to collect, catalogue and preserve all media publications issued in Germany, in German or about Germany since 1913. It also prepares and publishes the German National Bibliography, organised in seven series covering monographs, periodicals inside and outside the publishers' book trade, maps, university publications, music, and online publications.
The DNB holds a total of 49.6 million media. Of these, 35 million are physical media, including 18 million monographs and 9 million periodicals. A fast-growing number of online publications has amounted to 15 million items. In total more than 10,000 items are received each day.
Data available at the DNB can be accessed in various ways and formats. Data formats include MARC 21, Dublin Core, RDF for Linked Data, METS/MODS, CSV and PDF for the German National Bibliography and New Release Service.
Users can search in DNB holdings through the catalogue and export metadata lists via the data shop included in the catalogue. They can also submit queries to an interface such as SRU or OAI-PMH to obtain larger quantities of metadata, predefined selections of metadata or updates. Complete dump files of DNB holdings can also be downloaded from the website.[3]
A collection of freely available full text corpora is offered via the DNBLab. It also provides exemplary data analyses and tutorials for using the DNB holdings for various kinds of applications.

2 Deutsche National Bibliothek. https://www.dnb.de/EN/.

3 Deutche National Bibliothek. Metadata Services. https://www.dnb.de/EN/Professionell/Metadatendienste/metadatendienste_node.html.

The initiative to integrate DNB data into the GoTriple platform involved several preliminary steps. Firstly, DNB data was not available in a format for which GoTriple already provided import procedures. GoTriple ingestion pipelines support OAI-PMH with Dublin Core and the Europeana Data Model, as well as bulk imports of OpenAIRE-compliant data dumps; however, none of these options are directly supported by DNB, which publishes metadata, among other formats, in MARC 21 format.

The latter therefore appeared to be the most promising approach. MARC 21 is a metadata format that is widely accepted and used by libraries in Germany (cf. Arbeitsgruppe Kooperative Verbundanwendungen 2021) and internationally and serves as the basis for creating a new import process via a data format connector. This also supports subsequent use of the process by other European libraries.

As a first step, the complex structures and rules of the MARC 21 format had to be analysed and subsequently aligned with the GoTriple internal data model.

As a second prerequisite, DNB holdings had to be filtered to obtain a collection of all Social Sciences and Humanities content. This was achieved using the DDC Subject Categories - a classification system based on the Dewey Decimal Classification (DDC), which provides a broad assignment of publications to subject areas in approximately 100 subject categories. The DDC Subject Categories also serve as the structure for the German National Bibliography.

3. GoTriple implementation of MARC 21 Connector

GoTriple[4] is a multilingual discovery platform for the Social Sciences and Humanities (SSH). It is the main outcome of the TRIPLE EU funded research project which ended in March 2023. GoTriple can be considered as a "vertical" search engine that indexes metadata of SSH-related assets, particularly documents (articles, theses, publications, datasets, etc...), projects and author profiles. At the time of writing, GoTriple presents over 20 million documents, about 26,700 projects, and 18.1 million author profiles.

Multilingualism is one of its main features and it concerns both metadata processing and the localisation of its user interface. Users can not only find documents in the main European languages but also in Arabic and Turkish. To facilitate access to research outcomes, the title and abstract of an article are automatically translated into English when descriptions in this language are missing. Finally, users can access and use GoTriple in 12 European languages, including English, Croatian, German, and Italian.

GoTriple's content is populated through a software component called SCRE (Semantic Content and Retrieval Engine), which manages the retrieval, processing, and storage of document, project, and author metadata in the platform's indexes. It is a configurable platform operated by specifying rules through an easy-to-use web dashboard. These rules define how to acquire and process content and the frequency of retrieval.

Content processing in SCRE is managed by two elements: *Sources* and *Flows*. Each *Source* takes care of data acquisition and processing from a single point of origin (e.g. an OAI-PMH Endpoint or a files archive). *Flows,* on the other hand manage the publication of data in the GoTriple index.

[4] GoTriple. https://gotriple.eu.

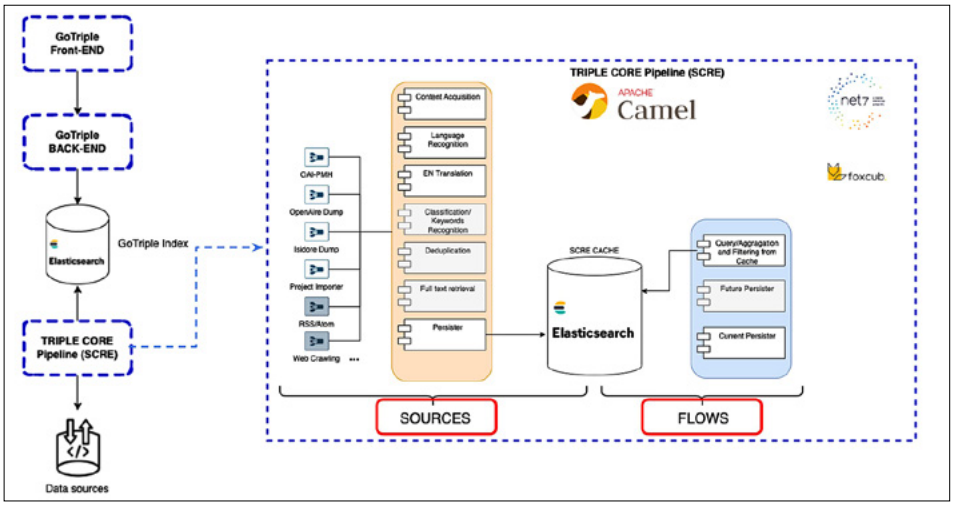

Figure 1. *SCRE software architecture*

SCRE uses a mediation technology based on the Apache Camel open-source platform to implement the processing workflow in a pipeline approach, with a set of services each specialised to perform a single specific task. These services can be classified as:

- *Connectors* in charge of retrieving metadata about publications and projects from specific data sources
- *Processors*, which curate or enrich the original metadata
- *Persisters*, which finally saves the enriched metadata in the platform's indexes.

Connectors can be of two types, *Harvesters*, which regularly and automatically access data from external sources using their APIs, and *Scanners*, which access and process file dumps that, must first be manually copied to in the SCRE local file system.
The fundamental hypothesis to consider is that the fetched content relates to the SSH domain. In fact, SCRE cannot filter out document or project metadata that is not relevant for GoTriple.
Connectors are designed to support specific harvesting protocols and data formats. OAI-PMH is the most commonly used connector, as many aggregators and data providers support this standard protocol. In this case, content retrieval occurs continuously, typically by querying for updates each week. File dumps, on the other hand must be manually fetched from the content providers' repositories and copied into the SCRE file system. The dedicated service recursively accesses each provided dump file, checks its syntax and imports the content.
Connectors process publication and project metadata one item at a time: the metadata of a single publication or project is then aptly mapped to the TRIPLE data model. Mapping of DNB's MARCXML content, an XML serialisation of MARC 21, to the TRIPLE data model was the first step taken by the joint team working on this project. Specific mapping rules were defined after collaborative work in which various options were evaluated, taking into account the actual MARC 21 data from DNB.
Once the mapping was completed, the strategy for managing content retrieval from DNB

indexes was assessed. Several options were considered, starting with the possibility of implementing of the harvester-type connector, which significantly reduces the effort required to update the GoTriple index.

Initially, the use of OAI-PMH was investigated, as this protocol appeared to be the most promising option, due to its well-established nature and wide support among data sources, including libraries. Although an initial prototype was implemented, this option had to be abandoned because the DNB OAI-PMH endpoint enforces a hard upper limit of 100,000 records per harvesting session. While records are delivered in paginated batches via resumption tokens, the total number of retrievable documents cannot exceed this threshold. This limitation is significantly below the scale required for the GoTriple integration, which aimed to ingest approximately 1.7 million DNB records, rendering the OAI-PMH approach unsuitable for this use case.

The second option considered was the development of a dedicated connector based on the SRU APIs that DNB provides to retrieve its content. SRU, which stands for "Search/Retrieval via URL", is a standard search protocol developed by the Library of Congress (LOC 2016) that allows advanced search and filtering capabilities. A first prototype was implemented but development was abandoned after a while because the DNB SRU API does not allow results to be filtered or ordered by the date of publication attribute. This limits SCRE's ability to perform incremental updates.

The only remaining option was therefore the implementation of a dedicated scanner connector, which processes export file dumps in MARCXML format regularly produced by DNB.

In this case, several issues needed consideration. First, it was necessary to accept content suitable for GoTriple, as DNB dump files include assets of various topics and types. It is in fact essential to import: (1) only those related to the SSH domain; and (2) those with a type compatible with the definition of "Documents" accepted by GoTriple, as defined in the Content Types controlled vocabulary of the TRIPLE ontology (TRIPLE Ontology 2025).

Also, DNB can only provide full dumps of their indexes: to perform an incremental update it is necessary to identify assets modified after a certain timestamp.

For all these reasons, a pre-processing script has also been developed. By taking as input the complete DNB data export in MARC 21, consisting of five separate files in XML format, the script analyses each record contained in them, discards those not suitable for GoTriple and saves in a separate file the metadata of the selected assets. Selection considers the type of the document (e.g. by analysing the data field tag="336" MARC 21 element) and its DDC subject category, to accept only those to be associated to the SSH domain (e.g. 4* - Language -, 8* - Literature -, 330 - Economics, ...). For subsequent iterations, the pre-processing script will extract only the documents that have been updated after a specified timestamp, ensuring that only those changed since the last import are considered.

From approximately 30 million records available in the DNB data dumps, a final selection of about 1.7 million assets was identified and subsequently ingested using a dedicated scanner connector developed for the project. This connector reused the xbib open-source library (https://github.com/xbib/marc) to facilitate the processing of MARC 21 metadata. The planned frequency for harvesting new content has been set to every six months, as is customary for file dump imports in GoTriple and aligned with DNB's biannual updates.

4. Challenges and future plans

From an external perspective, it may be surprising that exchange formats and more importantly metadata schemas can be very specific to a community and are used according to additional internal rules and procedures. The entire team had the task to learn about and familiarising with the DNB data catalogue and the corresponding data format MARC 21, which is currently the most widely used exchange format in the library context (alongside RDF).
Furthermore, we learned together what lies behind the millions of DNB entries, how to use the DDC subject categories and which types of entries are relevant for the community. The issues of data size and interface optimisation also had to be considered and may remain a challenge for future plans to integrate more data from national libraries into GoTriple. For this, rules for the selection process and harvesting needed to be defined to provide a standardised procedure with best practice characteristics, based on expert knowledge.
Probably, one of the biggest challenges for all data integration is data curation, even if you define rules for data mapping and implement them correctly. There will always be problematic individual cases in the metadata that inevitably cause confusion. Some of these problems cannot be solved or are retained as a compromise in the hope of either future improvement or with awareness of semantic incompatibilities. The most severe problems could be resolved by the developers of GoTriple in close cooperation with the data curation team. There is no way to manually check 1.7 million of new entries in GoTriple but even exemplary checks proved very useful for detecting some general problems that could then be addressed afterwards. One of the main and unique features of the European discovery platform is its multilingual options, including even automatic translation at the title level (see above). This proved problematic for bibliographic entries from a national library which are mostly in one language (in this case German) and include a certain number of translations that display the original title plus the German translation in the title field[5]. Adding on automatic translation based on users' browser language settings would inevitably cause confusion. Therefore, GoTriple changed this feature to a language switch button which users must press to see the English translation.
Another, more general problem is that the licence information in the metadata of each entry is extremely important for a European platform like GoTriple which is also an OPERAS service. In the long history of libraries, "open access" or "full text availability" is a relatively recent label included in the metadata of a bibliographical record. This results in a large number of "undefined" entries regarding the licence information and very few with the "full text available" label. Even if full text is available for several entries found, not every data provider (for example small institutional repositories represented in GoTriple via BASE) includes this metadata field in their schema, or they may leave it empty. The solution lies in the collaboration between libraries, repository providers and the European infrastructures behind discovery platforms like GoTriple.
Future plans could include integrating data repositories and/or the integration of not only descriptions of books, but also book parts, such as individual chapters. Additionally, possible extending GoTriple to other disciplines might be considered as other communities are interested in easily implementing their data as well.

[5] For an example see https://gotriple.eu/documents/dnb_963284606.

5. Integrate your own data catalogue / more (national) libraries in GoTriple

The integration of DNB content in GoTriple is important for two main reasons. Firstly, it provided a valuable opportunity to collaborate with a prestigious institution such as the German National Library, and to acquire a large amount of data for GoTriple. Partnering with a German institution was especially significant, given the outstanding work of the German node of OPERAS in promoting GoTriple and other OPERAS services to the national SSH community (OPERAS-GER 2023). Secondly, the implementation of MARC 21 connectors facilitates the integration of new data providers into GoTriple, given the widespread adoption of this standard, particularly by national libraries.

While the DNB integration, as indicated here, required specific implementations, two generic MARC 21 connectors will be included in SCRE to allow the import of MARC 21 XML content either via the OAI-PMH protocol or by scanning file dumps. As mentioned, initial prototypes have already been developed, and while their complete release is scheduled for year 2026.

It is important to emphasise once again that, for integration into GoTriple, providers must publish and provide only content related to SSH and of the appropriate type.

Further information on the GoTriple harvesting process is available in two documents, the "GoTriple content providers handbook" (Gingold 2023) and the previously mentioned "For content providers: your data in GoTriple" (De Santis 2023), both accessible from the GoTriple website.

Conclusion

Firstly, the addition of 1.7 million new entries from the DNB to GoTriple strengthens the platform and supports the goal of data growth. In particular, GoTriple's multilingual and open access approach means that, each integration of national library resources is a significant improvement. The German National Library benefits from a wider interconnectedness and visibility of its data. Domain-specific communities can more easily discover relevant content provided by the DNB promoting closer cooperation and expanded networks between the library and the SSH research community at the European level.

We hope that more national libraries will come forward and make the effort to integrate their well-curated bibliographic data into GoTriple. At the same time, we will continue to work with European and national research infrastructure consortia such as OPERAS and Text+ to connect data, people and platforms across borders.

Acknowledgments

The authors wish to acknowledge Danilo Giacomi and Tony Agosta of Net7 srl for their contributions in developing the software components necessary to integrate DNB sources into GoTriple.

References

Arbeitsgruppe Kooperative Verbundanwendungen. 2021. "Vereinbarungen der Arbeitsgruppe kooperative Verbundanwendungen zum Datentausch in MARC 21". https://www.agkva.org/888242273.html.

Brünger-Weilandt, Sabine, Kai-Christian Bruhn, Alexandra W. Busch, Erhard Hinrichs, Wolfram Horstmann, Martin Grötschel, Johannes Paulmann, et al. 2019. "Memorandum of understanding by NFDI Initiatives from the humanities and cultural studies," July. https://doi.org/10.5281/zenodo.3265763.

De Santis, Luca 2023. "Your data in GoTriple." https://gotriple.eu/docs/gotriple-handbook-v1_0.pdf.

Dumouchel, Suzanne. 2022. "About maintenance and knowledge infrastructures: some thoughts and examples from OPERAS experience". https://doi.org/10.5281/zenodo.6782623.

Gingold, Arnaud, & Luca De Santis. 2023. GoTriple content providers handbook. https://gotriple.eu/docs/gotriple-handbook-v2_0.pdf.

Larkin, Brian. 2013. 'The politics and poetics of infrastructure'. *Annual review of anthropology* 42, 327–43. https://doi.org/10.1146/annurev-anthro-092412-155522.

Library of Congress (LOC). 2016. "SRU- Search/Retrieve via URL". https://www.loc.gov/standards/sru/.

OPERAS-GER. 2023. OPERAS Germany web page. https://operas-ger.hypotheses.org/.

TRIPLE Ontology 2025. TRIPLE ontologies: content types vocabulary. https://gotriple.eu/ontology/triple/ContentType.

Sažetak

Postupak integracije knjižničnih podataka u GoTriple: primjer suradnje s Text+ konzorcijem

Cilj. Ovaj rad predstavlja primjer suradnje između nacionalne i europske istraživačke podatkovne infrastrukture i usluge koja je namijenjena zajednici. Najprije su pojašnjeni organizacijski i tehnički pristupi, a kasnije su na primjeru suradnje sa sudionicima predloženi načini integracije podataka. Na kraju se raspravlja o mogućem prijenosu i primjeni predstavljenog tijeka rada na druge primjere.

Pristup/Metodologija. Nakon zajedničke analize koju su proveli timova Text+ i GoTriple, razvoj namjenskog MARC 21 konektora za SCRE prepoznato je kao optimalna strategija za integraciju sustava Text+ u GoTriple. To je omogućilo uvoz publikacija iz područja društvenih i humanističkih (SSH) znanosti sadržanih u katalogu Njemačke nacionalne knjižnice (DNB) u GoTriple. Ova je odluka također olakšala integraciju relevantnih zbirki SSH-a kojima upravljaju druge knjižnice, budući da je MARC 21 važan standard za katalogizaciju i razmjenu podataka.

Rezultati. U ovoj integraciji bilo je nekoliko tehničkih izazova. Prije svega, MARC 21 je složen protokol koji se ne obrađuje lako. U ovom je slučaju upotreba namjenskih softverskih knjižnica otvorenog koda bila osobito korisna. Također se moralo procijeniti i testirati više mogućnosti prije nego što je pronađeno prikladno rješenje za prikupljanje podataka iz DNB-a u GoTriple.

Originalnost/Vrijednost. Predstavljeni primjer platforme GoTriple i jednog od izvora podataka Text+ (Njemačke nacionalne knjižnice) može poslužiti kao najbolja praksa za integraciju većeg broja podataka iz knjižničnih kataloga u platformu GoTriple. Također pruža model za učinkovitu komunikaciju i primjer suradnje potrebne za procjenu uloženog napora za buduću integraciju podataka u GoTriple na razini nacionalne infrastrukture.

KLJUČNE RIJEČI: bibliografski podaci, društvene i humanističke znanosti, GoTriple, integracija podataka, knjižnični katalog, metapodaci, OPERAS istraživačka infrastruktura, otkrivanje podataka, Text+