

Annif and Finto AI: Developing and Implementing Automated Subject Indexing

Osma Suominen^(a), Juho Inkinen^(b), Mona Lehtinen^(c)

a) National Library of Finland, <http://orcid.org/0000-0003-0042-0745>

b) National Library of Finland, <http://orcid.org/0000-0002-6497-6171>

c) National Library of Finland, <http://orcid.org/0000-0002-4735-0214>

Contact: Osma Suominen, osma.suominen@helsinki.fi; Juho Inkinen, juho.inkinen@helsinki.fi;
Mona Lehtinen, mona.lehtinen@helsinki.fi

Received: 5 May 2021; **Accepted:** 21 May 2021; **First Published:** 15 January 2022

ABSTRACT

Manually indexing documents for subject-based access is a labour-intensive process that can be automated using AI technology. Algorithms for text classification must be trained and tested with examples of indexed documents, which can be obtained from existing bibliographic databases and digital collections.

The National Library of Finland has created Annif, an open source toolkit for automated subject indexing and classification. Annif is multilingual, independent of the indexing vocabulary, and modular. It integrates many text classification algorithms, including Maui, fastText, Omikuji, and a neural network model based on TensorFlow. Best results can often be obtained by combining several algorithms. Many document corpora have been used for training and evaluating Annif. Finding the algorithms and configurations that give the best quality is an ongoing effort.

In May 2020, we launched Finto AI, a service for automated subject indexing based on Annif. It provides a simple Web form for obtaining subject suggestions for text. The functionality is also available as a REST API. Many document repositories and the cataloguing system for electronic publications at the National Library of Finland are using it to integrate semi-automated subject indexing into their metadata workflows. In the future, we are going to extend Annif with more algorithms and new functionality, and to integrate Finto AI with other metadata management workflows.

KEYWORDS

Automated subject indexing; Artificial intelligence; Machine learning; Metadata.

Introduction

Extensive digitization of paper archives and more active archiving of digital material are creating growing collections of data. Subject indexing, i.e. assigning documents with subjects from a controlled vocabulary, is an important method of organizing collections and improving their discoverability. Traditionally, subject indexing is a manual process performed by human experts, but since manual indexing is a very labour-intensive process, automated and semi-automated methods for subject indexing have been developed since the 1960s (Stevens 1965).

Automating some of the subject indexing processes in Finnish libraries and related institutions has long been a goal of the National Library of Finland for several reasons: to reduce the amount of indexing work, to make the subject indexing more consistent, and to expand subject indexing to collections where traditional manual indexing is not feasible. However, from our perspective, the existing tools and services for automated subject indexing suffer from a number of problems. First, our national languages, Finnish and Swedish, are not well supported by most tools. Second, the tools often rely on their own vocabulary, while we would like to use the General Finnish Ontology YSO¹ (Niininen, Nykyri, and Suominen 2017) as well as other Finnish subject vocabularies. Third, many of the available solutions are commercial services where the customer has little control of the system and is subject to vendor lock-in.

We started the development of Annif², our own open source tool for automated subject indexing, in 2017. Three years later, in May 2020, we launched Finto AI – an Annif based automated subject indexing service intended for production use³. In this paper, we explain the process of developing Annif, the text classification algorithms it supports, the quality assurance process we use to ensure that the algorithmically produced subject indexing meets expectations, the systems where Annif or Finto AI based automated subject indexing has been deployed, and conclude with some lessons learned.

Development of Annif

The first prototype of Annif was created in 2017, in an experiment to see if it was possible to use freely available metadata from the Finna⁴ discovery system to assist in the generation of new metadata (Suominen 2019). After a successful demonstration of the approach, the National Library of Finland decided in 2018 to start the development of a new version of Annif built on a more solid technical foundation and a set of goals and principles:

1. The tool should be multilingual, because in Finnish libraries, there is a need to support at least three languages: the national languages Finnish and Swedish, as well as English.
2. The tool should be independent of the indexing vocabulary; although the General Finnish Ontology is the most commonly used vocabulary in Finnish libraries, other special purpose vocabularies and library classifications such as the Dewey-based Public Library Classification YKL are widely used as well.

¹ <https://finto.fi/yso/en/>

² <https://annif.org/>

³ <https://ai.finto.fi/>

⁴ <https://finna.fi>

3. The tool should support different subject indexing algorithms; a general framework that can accommodate different algorithms was seen as more flexible and adaptable to different situations.
4. The tool should have a command line interface, a web user interface, and a REST API suitable for integration with other systems.
5. The tool should be provided as community oriented open source software; the National Library of Finland advocates for the use of open source software, as part of general openness and transparency goals, and the Skosmos⁵ vocabulary publishing software is following a similar open development model.

Based on the above goals, we created a modular architecture for Annif (Figure 1). User interaction is handled either through the command line interface (CLI) or the REST-style API that can be used to integrate Annif with other metadata management systems; the Finto AI web user interface, shown on the left in Figure 1, is an example of such a system. An embedded web user interface that relies on the REST API can also be used for interactive testing.

The *evaluation module* handles the calculation of various evaluation metrics such as precision, recall and F1 score. Annif is configured using a configuration file, handled by the *configuration module*. The *analyzer modules* support tokenization and normalization (stemming or lemmatization) of many languages. Indexing vocabularies, in either SKOS or a simple text format, are handled by the *vocabulary module*. The subject indexing algorithms are implemented as *backends*. The basic unit of configuration is a *project*, which is defined by specifying an indexing vocabulary, language, analyzer, backend, and project- or backend-specific parameters.

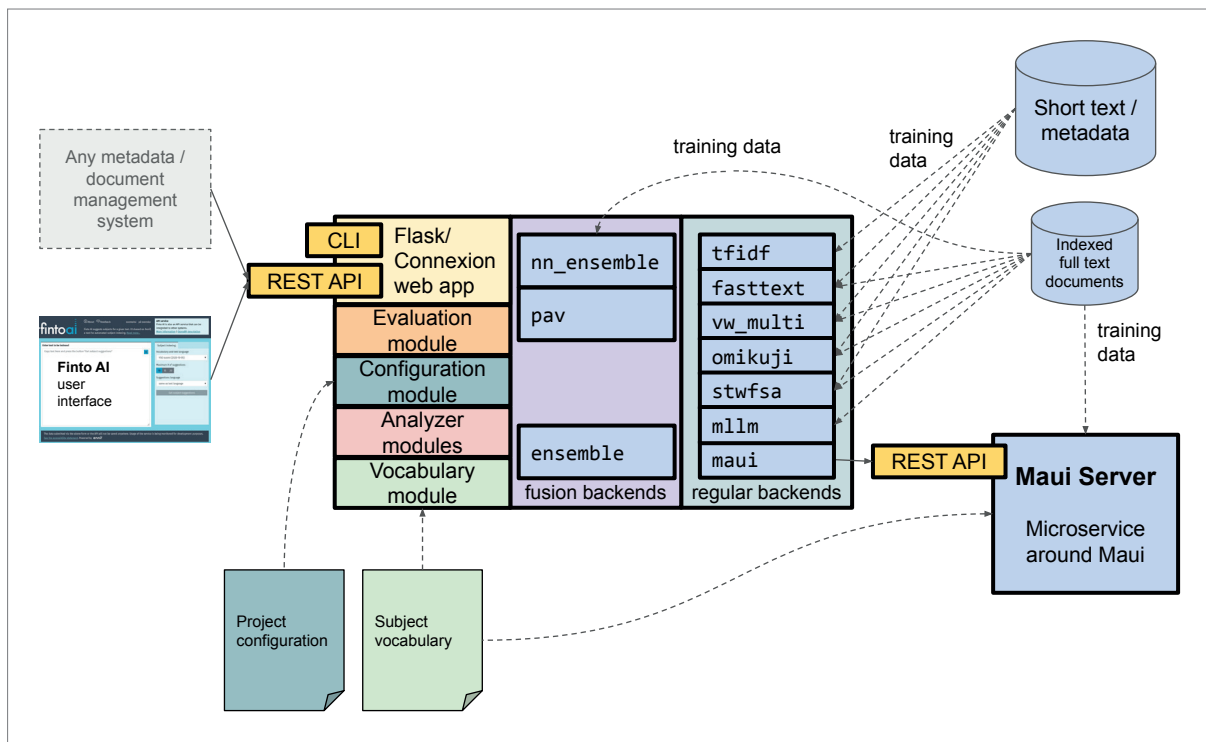


Fig. 1. Modular architecture of Annif

⁵ <https://skosmos.org>

Currently all the development of Annif happens on GitHub⁶. Annif is also made available as a Python package⁷ and as Docker images⁸.

Algorithms in Annif

Annif includes support for several text classification algorithms and thanks to the modular architecture, more can be added over time as backends. Backends can either function as *regular backends* or *fusion backends*. Regular backends work directly on document text and produce a suggestion of possible subjects. The algorithms implemented as regular backends in Annif are based on two main approaches: *lexical approaches* and *associative approaches* (for the distinction, see Toepfer and Seifert 2020). Fusion backends, also called *ensemble* backends, instead use the suggestions from other backends as input and produce a combined suggestion. Backends can thus be stacked and combined in many different ways.

Lexical approaches

In the lexical approach, words within document text are matched with the terms contained in the subject vocabulary. For example, if the vocabulary includes the term *gross national product* and its abbreviation *GNP* (for example as an alternate label for the same concept), then that concept will be suggested as a potential subject for a document containing either the full term or the abbreviation. Since a long document will contain many such matches, lexical algorithms also need to filter and select the most promising candidates; this is typically implemented using heuristics and machine learning.

Maui (Medelyan 2009) is an example of a lexical algorithm, and is supported in Annif by integration with Maui Server⁹. STWFSa (Toepfer and Seifert 2020) is another lexical algorithm supported in Annif by integration with its Python implementation¹⁰. It has been designed specifically for extracting the maximum information from short text such as metadata records for academic publications. We have created MLLM¹¹ (Maui-like Lexical Matching), a Python reimplementation of many of the ideas in the Maui algorithm, with some adjustments such as a different string matching method and new heuristics. All the previously mentioned lexical algorithms must be trained with a sample of manually indexed documents.

Associative approaches

In the associative approach, a statistical or machine learning model is trained on a large number (typically hundreds of thousands or more) of manually indexed documents in order to find words

⁶ <https://github.com/NatLibFi/Annif>

⁷ <https://pypi.org/project/annif/>

⁸ <https://quay.io/repository/natlibfi/annif>

⁹ <https://github.com/TopQuadrant/MauiServer>

¹⁰ <https://github.com/zbw/stwfsapy>

¹¹ <https://github.com/NatLibFi/Annif/wiki/Backend:-MLLM>

or expressions that correlate with particular subjects. For example, the subject *renewable energy sources* could be correlated with expressions such as “energy”, “solar power”, “fossil free”, “zero carbon”, “smart grids” and “battery technology” that appear frequently in documents indexed with that subject, even though not all of them are strictly related to the subject and may not appear at all as terms in the indexing vocabulary. When a well trained associative algorithm is given a new document containing such expressions, it is likely to suggest that it could be about renewable energy sources.

As a baseline method, Annif provides a simple associative backend called TFIDF that calculates a vector representation for each subject based on the words that appear in documents about that subject. When given a new document, the model suggests the most similar subjects for the words in that document, based on vector similarity. *fastText* (Joulin et al. 2016) is a fast and versatile machine learning algorithm for text classification created at Facebook Research and is supported in Annif by integration through its Python bindings. *Vowpal Wabbit* (VW) is a general purpose online machine learning framework; Annif supports its algorithms for multi-class and multi-label classification, which are generally best suited for relatively small vocabularies. Finally, *Omi-kuji*¹² is a reimplementation of a family of efficient tree-based machine learning algorithms for multi-label classification, including *Parabel* (Prabhu et al. 2018) and *Bonsai* (Khandagale, Xiao, and Babbar 2020); it is currently the most versatile and generally best performing associative algorithm in Annif.

Fusion approaches

A fusion approach, i.e. combining different kinds of automated subject indexing algorithms, can be an effective way of improving overall performance (Toepfer and Seifert 2020). Annif provides three fusion backends: a simple ensemble backend, which calculates a weighted average of suggestions from several sources; and two more advanced ensemble backends which require separate training with collections of manually indexed documents. The PAV ensemble (Pool Adjacent Violations) uses *isotonic regression* to estimate probabilities of particular subject suggestions being correct, based on the documents the ensemble has been trained on (see Wilbur and Kim 2014), and combines the estimated probabilities to calculate an overall suggestion. The TensorFlow based neural network ensemble combines the simple averaging method of the simple ensemble with a multi-layer perceptron network that learns how to adjust the combined suggestions so that they best match the manual indexing that the ensemble was trained on.

Quality of automated subject indexing

As we have developed tools and services for automated subject indexing, we have assessed the quality of the automated subject indexing process along the way. According to the framework presented by Golub et al. (2016), the quality of automated subject indexing can be approached from multiple perspectives:

¹² <https://github.com/tomtung/omikuji>

1. Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard.
2. Evaluating indexing quality in the context of an indexing workflow.
3. Evaluating indexing quality indirectly through retrieval performance.

We have so far focused on the first two perspectives, as the retrieval systems affected by the automated subject indexing processes (e.g. Finna) are quite far removed from the subject indexing processes and affected by numerous other factors as well.

API service configurations to evaluate

While we have performed many evaluations of individual algorithms during the development of Annif, the most thorough evaluations have been performed on the combinations of projects, backends, configuration settings, and training data sets that have been provided for public use in the API service for Annif and (since May 2020) Finto AI. The first public API service, after the initial prototype, was published in January 2018, with support for suggesting subjects from the General Finnish Ontology YSO for documents in Finnish, Swedish or English. We set up an ensemble project combining results from three different algorithms for each language. The associative algorithms were trained using metadata extracted from the Finna discovery system, while lexical and ensemble backends were trained on various collections of full text documents. Subsequently we have updated the API service with newer versions of the YSO vocabulary (including YSO Places from January 2020 onwards) and switched the backend algorithms and the ensemble type as new options have been developed. The changes to the API service configurations have been summarized in Table 1.

Date	YSO version	Ensemble type	Backends
2018-01	2017-03 snapshot	Simple ensemble	TFIDF, fastText, Maui
2020-01	2019-03 Cicero	Simple ensemble	Omikuji-Parabel, Omikuji-Bonsai, Maui
2020-03	2020-01 Diotima	Neural network	Omikuji-Parabel, Omikuji-Bonsai, Maui
2020-12	2020-10 snapshot	Neural network	fastText, Omikuji-Bonsai, Maui
2021-04	2021-03 Epikuros	Neural network	fastText, Omikuji-Bonsai, MLLM

Table 1. API service configurations.

Comparison to gold standard

A gold standard is a collection in which each document is assigned a set of subjects that is assumed to be complete and correct (Golub et al. 2016). Once a gold standard has been developed, it is easy to evaluate automated subject indexing methods against it by measuring how well the algorithmic suggestions match the gold standard. However, creating a good quality gold standard takes a significant amount of effort and requires input from many experts. In practice, existing manually indexed documents are often used as a substitute for a properly constructed gold standard, as in

the evaluation of the Maui algorithm (Medelyan 2009). Such collections are readily available and they enable easy experimentation and comparison of different algorithms, but as the indexing process is susceptible to many kinds of bias, they are best used as ballpark estimates of quality and must be complemented with other types of evaluation.

We have used the following manually indexed corpora for evaluation. The first three include documents in Finnish, Swedish and English, the last two are only in Finnish.

1. JYU theses: Master's and doctoral theses from the University of Jyväskylä (n=7,400) published in the years 2010 to 2017. These are long, in-depth academic documents that cover many disciplines.
2. Electronic deposits: Non-fiction electronic books (n=9832) published between 1998 and 2019 that have been deposited to the National Library of Finland and indexed in the national bibliography Fennica.
3. Book descriptions: Titles and short descriptions of non-fiction books (n=51309) collected from the database of the book distributor Kirjavälitys Oy, covering the time period from approximately 2000 to 2019. The book descriptions were originally created by publishers for marketing purposes. The subject indexing for these works was obtained separately from the national bibliography Fennica.
4. Ask a Librarian: Question and answer pairs from the Ask a Librarian service run by public libraries in Finland. The original database consisted of over 25,000 documents but we extracted the subset with a minimum of 4 subjects per document (n=3,150). These are short, informal questions and answers about many different topics.
5. Satakunnan Kansa: Digital archives of Satakunnan Kansa regional newspaper. The archives consist of over 100,000 unindexed documents. Out of these, a random sample of 50 documents was manually indexed by four librarians working independently.

We split these collections into train, validate and test subsets. Only the test subsets were used as gold standard sets for the evaluation of algorithms. We mainly used the F1@5 metric for the evaluation: that is, the F1 score similarity (harmonic mean of precision and recall) between the manually assigned subjects and the top 5 suggestions of the algorithm. The results are summarized in Figure 2. We can see that the overall F1 scores have generally improved with successive API service configurations. The best F1 scores of around 0.6-0.7 were obtained with Swedish language documents from the JYU theses and electronic deposit collections; however, these measurements are also the least reliable, since due to the small number of Swedish language documents in these collections, we had reused some of the same documents for both training and evaluating the Maui, MLLM and neural network ensemble models. If we exclude the two Swedish language collections with unrealistically good results, we have reached F1 scores ranging between 0.3 and 0.5.

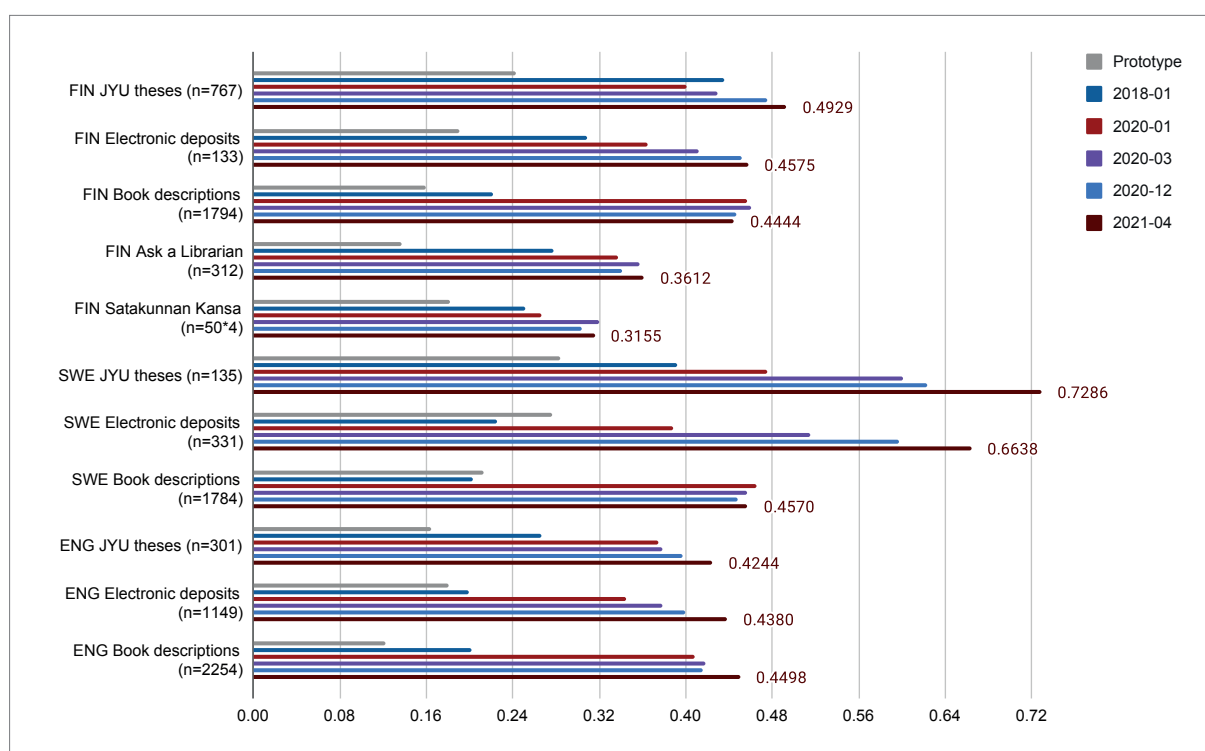


Fig. 2. F1@5 scores for the test collections, by API service configuration. The most recent numeric scores for the 2021-04 API service configuration are also shown

Assessment by evaluators

Having human evaluators assess the suggested subjects is another way to measure the quality of automatic subject indexing. In 2019, we organized a workshop where 48 participants (mainly librarians and informaticians) were given 50 example documents, with on average more than 10 sets of subjects assigned to each document. The indexing had been created either by humans (professional or lay) or by different Annif algorithms, but the participants did not know which was which. The participants used a scale from 1-5 to evaluate the indexing from three viewpoints: overall quality, meaningfulness and coverage. In general the human assigned subjects got higher scores, but the difference wasn't very large. Figure 3 shows the evaluation results. Indexing by the best performing Annif PAV ensemble model usually received a grade of around 3 out of 5, while human indexers scored between 3.5 and 4 out of 5, with professionals performing the best. Annif-assisted semi-automatic indexing landed in between. (Lehtinen, Inkinen, and Suominen 2019)

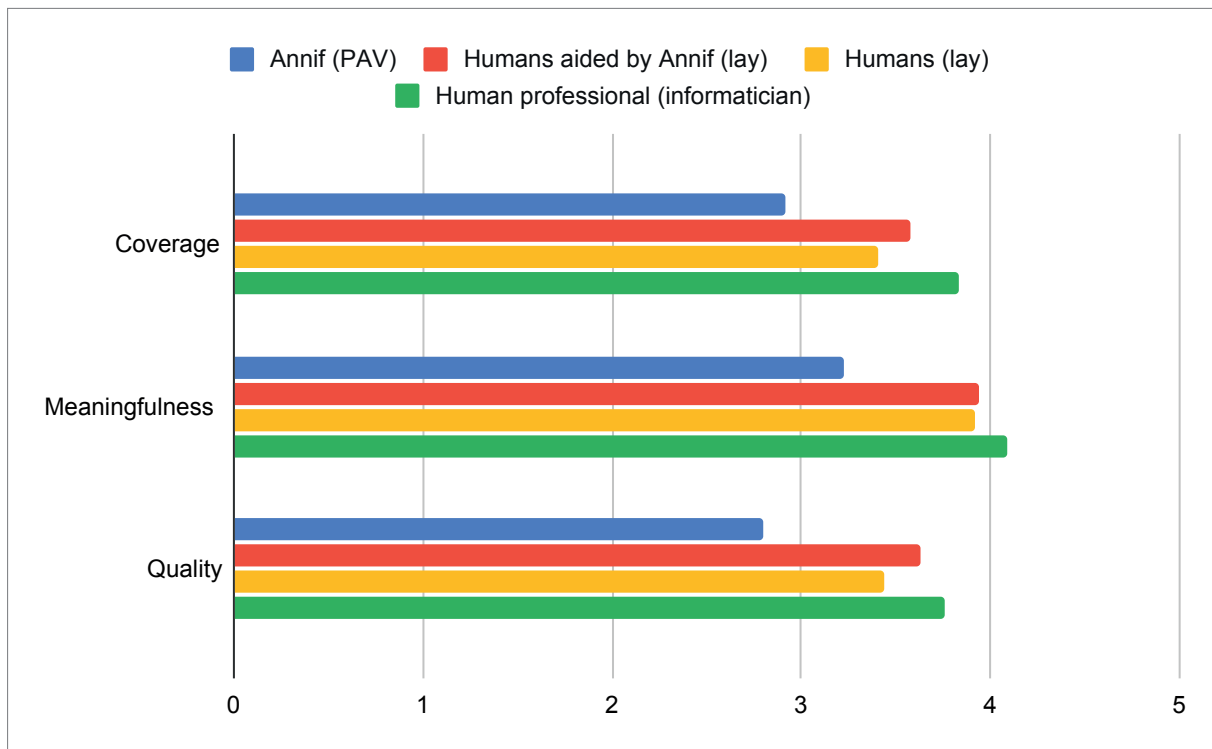


Fig. 3. Quality evaluation of intellectually given and Annif-produced subject indices. Data reproduced from Lehtinen, Inkinen, and Suominen 2019

A similar comparison was performed by the Finnish Public Broadcasting Company Yle. Their tests compared Annif against a commercial document classification service Leiki which they have been using in production for several years. In their results, Annif was rated as slightly better than Leiki for Finnish language documents and as much better for Swedish language documents. The quality of the metadata they used for training Annif might explain the differences between the languages (Suominen and Virtanen 2020; Nikkarinen 2021).

The Research department of the National Library of the Netherlands has also evaluated and used Annif as a part of developing their own larger tool for automated indexing (Haighton and Veldhoen 2020). The German National Library has evaluated Annif as well, comparing it with their current automated indexing system both qualitatively and quantitatively. Seven out of nine Annif's algorithms outperformed the current solution in F1@5 scores. Human evaluators also rated Annif's suggestions as more useful than those of the current system (Uhlmann 2020).

Evaluating in the context of an indexing workflow

We have also evaluated the quality of Annif in the context of the indexing workflow of the JYX¹³ institutional repository of the University of Jyväskylä, which was an early adopter of Annif. JYX

¹³ <https://jyx.jyu.fi/>

integrates Annif into its upload form. Students who upload their completed Master's thesis receive suggestions from Annif and can accept or reject the suggestions as well as add their own keywords. Later in the process, informaticians validate the metadata and can make corrections to the subjects. The system saves the original suggestions by Annif as well as the users' choices, so it is possible to keep track of how many of the Annif suggestions are accepted by the student and the final validated subjects. Figure 4 shows the F1 score similarities between the original Annif suggestions and the student-selected and final subjects, over several generations of API service configurations. There was a marked increase in the similarity after the initial prototype; since then, a small increase in similarity to the Annif suggestions can be seen in both the student-selected and final subjects, suggesting that the acceptability of the automated suggestions has increased over time.

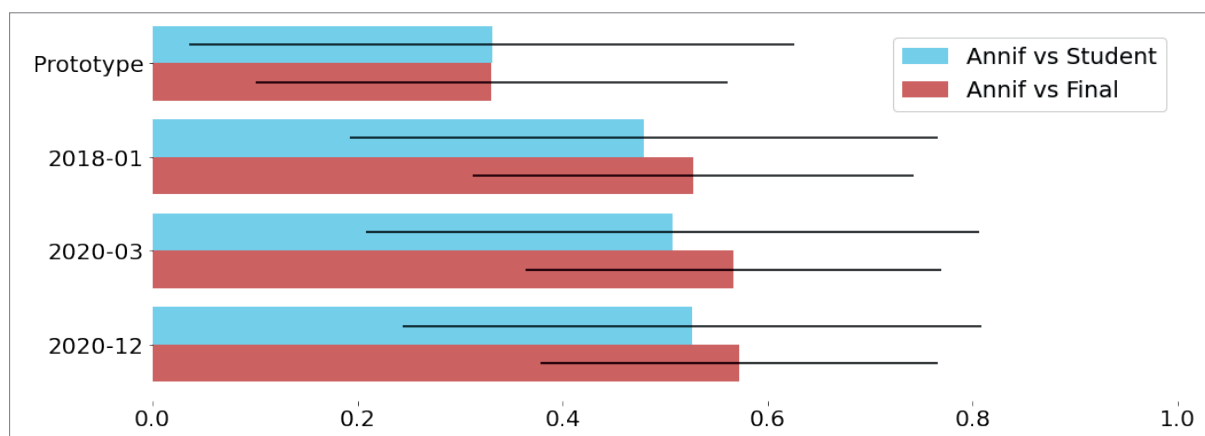


Fig. 4. F1 score similarity between Annif suggestions, student-selected subjects and final subjects in JYX, for the Annif prototype and subsequent API service configurations. Data is missing for the short-lived 2020-01 configuration

Users of the Annif API service and Finto AI

A service for automated subject indexing based on Annif has been existing since 2017 at the annif.org website, but its main purposes have been testing and development. The Finto AI service we launched in May 2020 is intended for production use. The service offers an easy way for introducing automatic subject indexing into information systems, provided that the vocabularies and language support offered by the API service meet local requirements. Some of the systems integrated with Finto AI are shown in Figure 5.

Generally, when the API service is integrated in the indexing workflow of a document repository, the steps in processing a document are:

1. extract the text from the document (typically a PDF file)
2. detect the language of the text (if not already known)
3. send the text to Annif via the *suggest* method of the API; the specific endpoint is chosen based on the text language and the indexing vocabulary
4. display the returned subject suggestions to the user
5. the user selects the subjects to be stored in the document metadata; the user can also add subjects that were not suggested by Annif

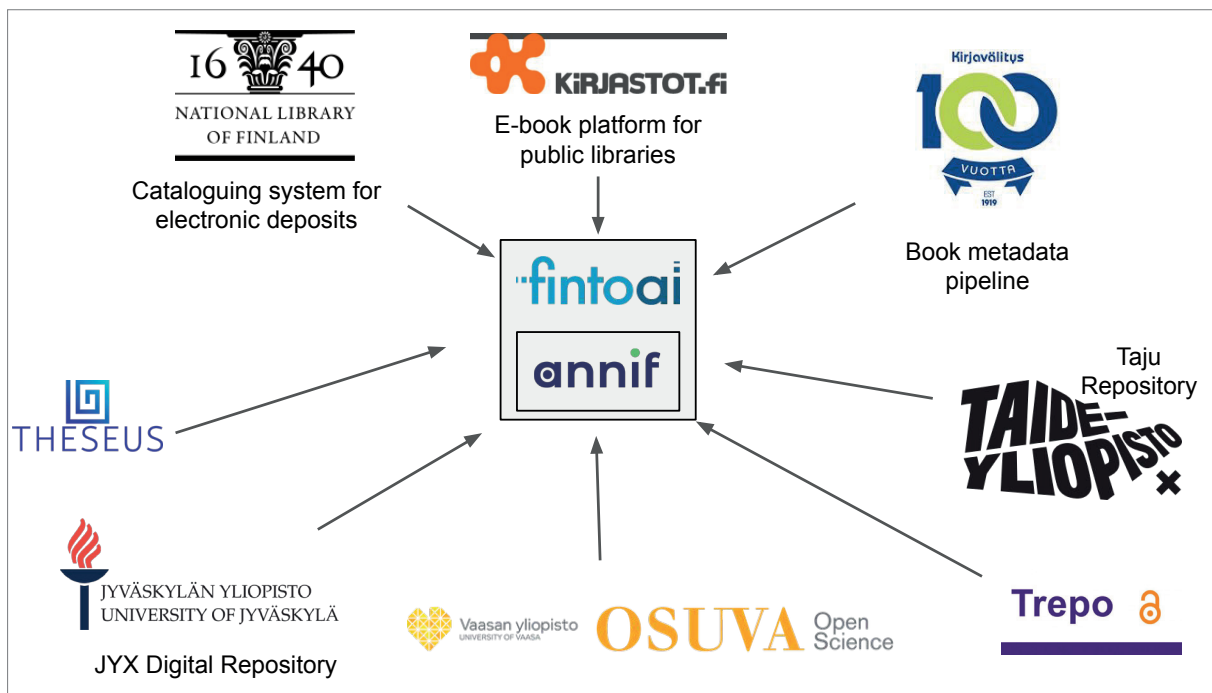


Fig. 5. Institutional users of Finto AI

Institutional repositories

The very first institutional user of semi-automated subject indexing by Annif was the JYX repository of the University of Jyväskylä, which is based on DSpace software. Already in 2017 they integrated the API of the Annif prototype system into their pipeline, which is used by students to upload their Master's or doctoral theses. As explained above, a librarian may correct the student-selected subjects when validating the metadata.

Since 2020, until April 2021 when this article was written, four DSpace based university repositories maintained by the National Library of Finland have started using Finto AI in their uploading pipeline: Osuva¹⁴ (University of Vaasa), Trepo¹⁵ (University of Tampere), Taju¹⁶ (University of Arts) and Theseus¹⁷ (used by many Finnish universities of applied sciences). Their workflow is similar to JYX.

The electronic deposit system at the National Library of Finland

The National Library of Finland maintains an uploading service for individual deposits of electronic publications¹⁸. The API of Finto AI was integrated in 2020 to the metadata workflow of the

¹⁴ <https://osuva.uwasa.fi/>

¹⁵ <https://trepo.tuni.fi/>

¹⁶ <https://taju.uniarts.fi/>

¹⁷ <https://www.theseus.fi/>

¹⁸ <https://luovutuslomake.kansalliskirjasto.fi>

internal deposit repository Varsta. The subject suggestions are not shown to the uploader, but to a library cataloguer who curates the metadata in the Varsta system. The metadata is then stored in the Melinda union catalogue. The publication files are stored in the Varia repository, which can be browsed using the computers within the premises of the National Library of Finland. See Figure 6 for an overview of the pipeline.

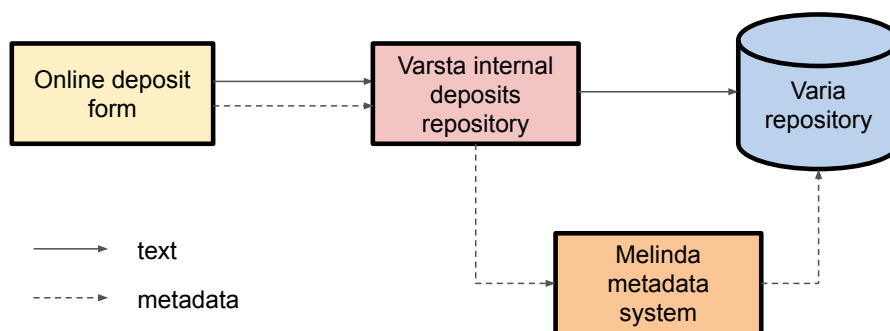


Fig. 6. Data flows for individual electronic publication deposits in the systems of the National Library of Finland

Book distributor Kirjavälitys Oy

Kirjavälitys Oy¹⁹ is a Finnish book distributor that handles book-sale logistics. They receive information about upcoming titles from publishers and produce metadata used by libraries, booksellers and the union catalogue Melinda, which includes the Finnish national bibliography Fennica (see Figure 7). Kirjavälitys has integrated the API of Finto AI in their system to aid in subject indexing of non-fiction books. They use the back-cover description text of books as the input to Finto AI.

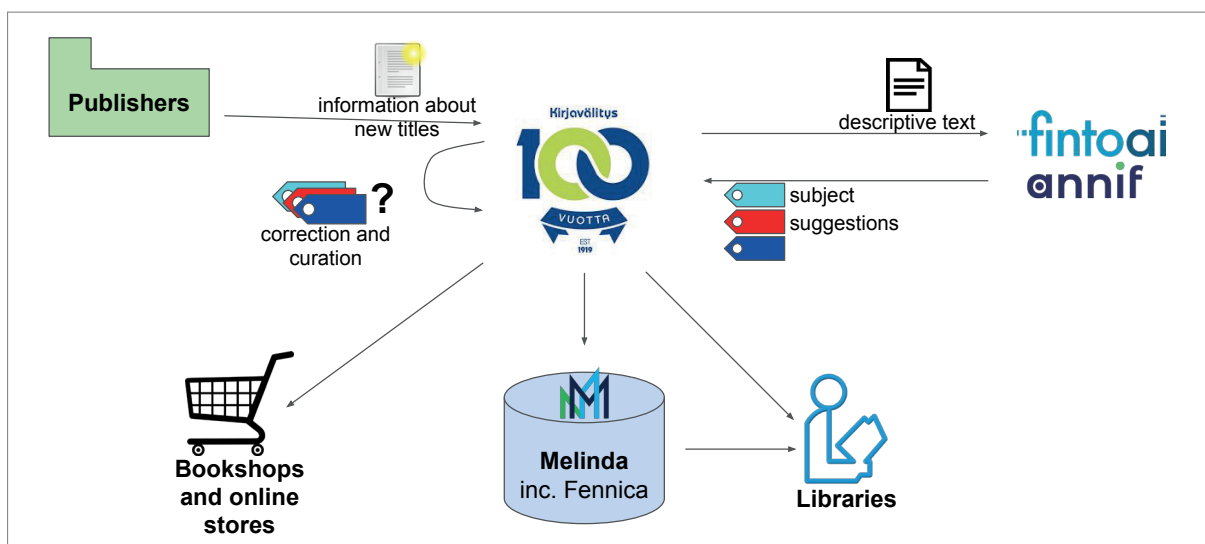


Fig. 7. The book distributor Kirjavälitys Oy receives information from publishers, enhances it with subject indexing assisted by Finto AI, and produces widely used metadata

¹⁹ <https://www.kirjavalitys.fi/en/home/>

Standalone Annif installations

To have more control on the indexing, e.g., for using a custom vocabulary or achieving better indexing quality on a specific topic area, or to support a language not available in Finto AI, a user can install and set up Annif by themselves and train their own models. Training a well-performing Annif model requires possessing adequate amounts of suitable training data, and can be computationally heavy. Searching for good hyperparameters for a model takes a lot of computation time. In contrast, when a model has been trained, and it is used by an Annif instance to offer subject indexing functionality via API, much less CPU resources are needed. For these reasons it can be worthwhile to have separate computing environments for training Annif models and for serving them.

Here we present some institutions that have set up their own Annif installations.

The Leibniz Information Centre for Economics ZBW has a long history of developing automated subject indexing solutions. Currently they are working on the AutoSE project with the aim of transferring their existing automation solutions into productive use (Kasprzik 2020). They use Annif as a part of their framework, and also actively contribute to the development of Annif.

The Finnish Broadcasting Company Yle is setting up Annif for semi-automatic subject indexing of online news articles. They use their own vocabulary and training corpus, and as their custom vocabulary evolves rapidly, they retrain their Annif models every week (Nikkari 2021).

The Finnish National Audiovisual Institute (KAVI) offers various services, such as film digitization, and maintains archives. They are also responsible for content rating and screenings of audiovisual material in Finland. KAVI first tested Annif as a standalone installation for indexing radio and TV programs using a speech-to-text transcript of the audio content. Based on the test results, KAVI decided to adopt Annif for this use in their future archive management system (Lehtonen and Piukkula 2020).

The National Library of the Netherlands has explored the possibilities of automatic indexing (Kleppe et al. 2019). Annif is now used as a part of their larger tool that is being developed for library cataloguers (Haighton and Veldhoen 2020). The training data for their current models have been gathered from a collaborative cataloguing system for Dutch libraries. The data consists of titles, subtitles and summaries of Dutch e-books. The Brinkman thesaurus²⁰ has been used as the controlled vocabulary. Annif has also been applied in a Dutch research project called Entangled Histories. The project focused on early modern ordinances, i.e. law texts, and Annif was used in their classification (Romein, Veldhoen, and de Gruijter 2020).

Dissemin²¹ is an online service for researchers to find open publishing repositories for their publications. Dissemin uses Annif to categorize academic pre- and postprints uploaded to open repositories.

²⁰ <https://www.kb.nl/sites/default/files/docs/brinkmanonderwerpen-2018.pdf>

²¹ <https://dissem.in/>

Community building

We aim to foster a community around Annif and to make it easy for people to learn about it. Annif has a website that serves as an introduction and an interactive demo. In the Annif GitHub project, we offer a thorough technical description and tips for Annif use. Users can also report bugs or contribute ideas and solutions using GitHub issues and pull requests. There is also a user forum called *annif-users*²² where people can ask for help, discuss and share their experiences. The forum is also a platform for Annif-related announcements and news.

Together with ZBW, we have created a hands-on tutorial²³ to help people get started with Annif. The first tutorial session was held at the SWIB19 conference²⁴. When the Covid-19 pandemic hit in 2020, we turned the material into an online tutorial suitable for self study, with videos on YouTube and exercises on GitHub. We have organized several interactive workshops based on the tutorial materials at suitable online conferences.

We also took part in the EU-funded High-Performance Digitisation project, which was a joint effort with CSC – IT Center for Science and the National Archives of Finland. The project sought to find intelligent solutions for automatic indexing workflow in LAM organizations. We were really pleased with this collaboration, which resulted in e.g. the discovery and thorough evaluation of the highly efficient Omikuji algorithms that were later integrated into Annif. The project is described on its web page²⁵ and in Lehtinen & Kallio (2020). The project also produced a whitepaper (in Finnish) describing the uses and challenges of automatic subject indexing in a cultural heritage organization, with Annif as an example (Hulkkonen et al. 2021).

Conclusion and Lessons Learned

Manually indexing documents for subject-based access is a labour-intensive process, and with the growing mass of digital material it becomes more and more difficult to keep up. There is a need for automation. Although it has taken several years and a lot of development effort, we have successfully created an open source solution for multilingual, vocabulary independent automated subject indexing that has become a production service used in many Finnish libraries, especially through the Finto AI service.

Annif is a unique framework into which different text classification algorithms can be integrated. The algorithms may be used alone, or in combinations called ensembles. We have found that the ensembles nearly always perform better than the individual algorithms.

Subject indexing is not an easy process, either for human indexers or for algorithms. Some parts of it are inherently subjective. When humans do subject indexing, they can have very different perspectives, or sometimes simply make mistakes. These types of mistakes or differences of opinion, however, are usually still relatable or understandable. When algorithms do subject indexing, their mistakes often do not necessarily make any sense from a human perspective.

²² <https://groups.google.com/g/annif-users>

²³ <https://github.com/NatLibFi/Annif-tutorial/>

²⁴ <https://swib.org/swib19/>

²⁵ <https://www.csc.fi/en/-/high-performance-digitisation>

There are many approaches for evaluating the quality of automated subject indexing systems. We have found that a combination of approaches works well for our purposes. Quantitative comparisons to a human indexed gold standard are the easiest to produce, and we perform them frequently both for the purpose of algorithm development and for evaluating the models that we deploy into production services. User oriented evaluation methods, such as assessment by evaluators, are more laborious, but they produce important insights about how algorithmically produced subject indexing differs from manually created indexing. Organizing workshops around automated subject indexing has provided a way of crowdsourcing the human evaluation effort, while simultaneously spreading awareness about automated indexing among librarians. We have also started to track how our tool is being used in the indexing workflow of systems that are using our API services. In the future, it would also be possible to investigate how the use of automated indexing affects users of retrieval systems.

The Annif tool is increasingly being deployed in Finnish library systems by integration with the API services provided by Finto AI. The Finto AI web user interface is also being used directly by librarians in cases where direct integration between systems is not feasible or has not yet been implemented. So far, users have been very positive towards the subject suggestions given by the service, as it provides an initial suggestion of potential subjects instead of an empty field to fill in. This is especially important for university library repositories where students, who are usually not experts in subject indexing, upload their own thesis documents.

The API services available through Finto AI are currently limited in the terms of indexing vocabularies and languages we can offer. We are working with Finnish organizations that have more diverse needs, for example custom domain-specific vocabularies, so that we can expand the service in the future.

Annif has been community oriented open source software from the start. We have created a web site and a wiki with technical documentation, set up a user forum, presented the tool at conferences and webinars, and together with ZBW, produced a tutorial for learning the basics of the tool. The effort put into community building is starting to pay off, as we are seeing an increasing number of test installations of Annif and some organisations are investing seriously in the adoption of Annif, for example by making extensive tests and comparisons.

One of the challenges in adopting Annif is collecting suitable training data and converting it to the corpus formats that Annif understands. This process usually requires programming skills. Even with a corpus in the correct format, achieving and maintaining good quality can be a challenge. We have gathered advice for setting up and refining projects into a wiki page²⁶.

There are upsides and downsides of the open source model for library systems. It allows for freedom and flexibility, but requires more technical expertise and resources than similar systems and services provided by commercial vendors. Organizations adopting an open source solution must be prepared to build the in-house expertise required to set up and maintain the systems. Some of the development effort can be shared and pooled through co-operating using code sharing platforms such as GitHub.

²⁶ <https://github.com/NatLibFi/Annif/wiki/Achieving-good-results>

In the future we continue to actively develop Annif and Finto AI. We hope to keep the community involved and welcome any contributions and feedback. Our aim is to support more vocabularies and languages in the Finto AI service while following the development of new text classification algorithms and utilizing them.

Acknowledgements

We thank the institutions and people who provided us with the corpora that have been used to train and evaluate the automated subject indexing methods, and Ari Häyrynen for providing the data used for the evaluation of Annif in the context of the JYX repository indexing workflow.

References

- Golub, Koraljka, Dagobert Soergel, George Buchanan, Douglas Tudhope, Marianne Lykke, and Debra Hiom. 2016. 'A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval'. *Journal of the Association for Information Science and Technology* 67 (1): 3–16. <https://doi.org/10.1002/asi.23600>.
- Haighton, Thomas, and Sara Veldhoen. 2020. 'Assisted Keyword Assignment Using Annif. KB Lab: The Hague.' 2020. <http://kbresearch.nl/annif/>.
- Hulkkonen, Juha, Juho Inkinen, Alekski Kallio, Markus Koskela, Mikko Lappalainen, Mona Lehtinen, Mats Sjöberg, Osma Suominen, and Laxmana Yetukuri. 2021. 'Sisällönkuvailun automatisoinnin haasteita ja ratkaisuja kulttuuriperintöorganisaatioissa'. Kansalliskirjaston raportteja ja selvityksiä. <http://urn.fi/URN:ISBN:978-951-51-7233-4>.
- Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. 'Bag of Tricks for Efficient Text Classification'. *ArXiv:1607.01759 [Cs]*, August. <http://arxiv.org/abs/1607.01759>.
- Kasprzik, Anna. 2020. 'Putting Research-Based Machine Learning Solutions for Subject Indexing into Practice'. In *Proceedings of the Conference on Digital Curation Technologies (Qurator 2020)*. Berlin, Germany. http://ceur-ws.org/Vol-2535/paper_1.pdf.
- Khandagale, Sujay, Han Xiao, and Rohit Babbar. 2020. 'Bonsai: Diverse and Shallow Trees for Extreme Multi-Label Classification'. *Machine Learning* 109 (11): 2099–2119. <https://doi.org/10.1007/s10994-020-05888-2>.
- Kleppe, Martijn, Sara Veldhoen, Meta van der Waal-Gentenaar, Brigitte den Oudsten, and Dorien Haagsma. 2019. 'Exploration possibilities Automated Generation of Metadata'. Zenodo. <https://doi.org/10.5281/zenodo.3375192>.
- Lehtinen, Mona, Juho Inkinen, and Osma Suominen. 2019. 'Aaveita koneessa: Automaattisen sisällönkuvailun arviointia Kirjastoverkkopäivillä 2019'. *Tietolinja* (blog). 2019. <http://urn.fi/URN:NBN:fi-fe2019120445612>.
- Lehtonen, Tommi, and Juha Piukkula. 2020. 'Automaattinen asiasanoitus Radio- ja televisio-ohjelmätietokanta Ritvassa'. *Informaatiotutkimus* 39 (1): 27–45–27–45. <https://doi.org/10.23978/inf.88107>.
- Medelyan, Olena. 2009. 'Human-Competitive Automatic Topic Indexing'. Thesis, The University of Waikato. <https://researchcommons.waikato.ac.nz/handle/10289/3513>.
- Niininen, Satu, Susanna Nykyri, and Osma Suominen. 2017. 'The Future of Metadata: Open, Linked, and Multilingual – the YSO Case'. *Journal of Documentation* 73 (3): 451–65. <https://doi.org/10.1108/JD-06-2016-0084>.
- Nikkarinen, Irene. 2021. 'Annif <3 Yle 2.0: Annifin osittainen käyttöönotto artikkeleiden koneavusteisessa asiasanoituksessa'. Presented at the Meeting of the Finnish Automatic Indexing Interest Group, March 15. <https://www.kiwi.fi/display/tekoalykumppanuus/Automaattisen+ku+vailun+verkoston+tapaamiset?preview=/147358597/211911484/Automaattisen%20kuvailun%20verkoston%20tapaaminen%2015.3.2021%20Annif.pdf>.

Prabhu, Yashoteja, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. 'Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising'. In *Proceedings of the 2018 World Wide Web Conference*, 993–1002. WWW '18. Lyon, France. <https://doi.org/10.1145/3178876.3185998>.

Romein, C. Annemieke, Sara Veldhoen, and Michel de Gruijter. 2020. 'The Datafication of Early Modern Ordinances'. *DH Benelux Journal* 2. <https://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html>.

Stevens, Mary Elizabeth. 1965. *Automatic Indexing: A State-of-the-Art Report*. NBS Monograph 91. Washington, D.C: United States. Government Printing Office.

Suominen, Osma. 2019. 'Annif: DIY Automated Subject Indexing Using Multiple Algorithms'. *LIBER Quarterly* 29 (1): 1. <https://doi.org/10.18352/lq.10285>.

Suominen, Osma, and Pia Virtanen. 2020. 'Yle Meets ANNIF – an Open Source Tool for Automated Subject Indexing'. Presented at the EBU MDN Workshop 2020, June 10. <https://tech.ebu.ch/contents/publications/events/presentations/mdn2020/yle-meets-annif--an-open-source-tool-for-automated-subject-indexing>.

Toepfer, Martin, and Christin Seifert. 2020. 'Fusion Architectures for Automatic Subject Indexing under Concept Drift: Analysis and Empirical Results on Short Texts'. *International Journal on Digital Libraries* 21 (2): 169–89. <https://doi.org/10.1007/s00799-018-0240-3>.

Uhlmann, Sandro. 2020. 'Automatische Vergabe von GND-Schlagwörtern Mit Annif - Ergebnisse Einer Evaluation Im DNB - Projekt EMa'. Presented at the Erfahrungen und Perspektiven mit dem Toolkit Annif, December 3. <https://wiki.dnb.de/display/FNMVE/Workshop+2020%3A+Toolkit+Annif>.

Wilbur, W. John, and Won Kim. 2014. 'Stochastic Gradient Descent and the Prediction of MeSH for PubMed Records'. *AMIA Annual Symposium Proceedings* 2014 (November): 1198–1207.

Towards an open and collaborative Authority Control

Barbara Katharina Fischer^(a)
with the cooperation of Jürgen Kett^(b),
Sarah Hartmann^(c), Mathias Manecke^(d)

a) Deutsche Nationalbibliothek (The German National Library)
b) Deutsche Nationalbibliothek (The German National Library)
c) Deutsche Nationalbibliothek (The German National Library)
d) Deutsche Nationalbibliothek (The German National Library)

Contact: Barbara Katharina Fischer, b.k.fischer@dnb.de
Received: 25 June 2021; **Accepted:** 23 July 2021; **First Published:** 15 January 2022

ABSTRACT

As digital transformation is speeding up, the need for a reliable retrieval is too. Libraries have long used *authority files* to enhance the search for information. Now, as the entire GLAM field is increasingly presenting its content online, national libraries face the requirement to provide authority data as reference points to a far more diverse community. The request is not limited to persistent identifiers but new records on non-librarian entities are needed. The German National Library (DNB) aims to provide an open framework that allows *collaboration* on all levels: editing the records, defining the regulations and standards plus ease the data flow in both directions. To this end, the DNB has started an ambitious project transferring both the authority file records and their regulations into a *Wikibase* instance. The article relates the findings working with the beta version of the software that drives *Wikidata*. To spur the process the DNB co-published the WikiLibrary Manifesto together with Wikimedia Deutschland. The institutions signing the manifesto shall cooperate to improve the building of a technical infrastructure that will ease knowledge equity through the *FAIR Data Principles* and the creation of a structured data ecosystem. The manifesto was signed by IFLA in June 2021.

KEYWORDS

Library; Wikibase; Authority control; Fair data; Semantic web.

“What really distinguishes us is the way in which we collaborate on a major scale.”¹

When people discuss topics with verve and persistence, this is generally a sign of dedication and connection. The topic of “Opening the GND” features these positive qualities. It affects and moves many people. It raises questions on the major topic of collaboration, both in great detail and a vast range of different contexts. The opening quote to this article is taken from the historian Yuval Noah Harari’s² much-acclaimed graphic novel “Sapiens”, which tells the history of how humankind developed. It also describes our work in the *Office for Library Standards (AfS)* at the German National Library. Organising collaborations is at the heart of what we do. Our task is to facilitate the cataloguing of knowledge resources across national and disciplinary boundaries. We organise collaborations by promoting consensus on standards that we ultimately use to describe the world while keeping them equally comprehensible for all. Using these standards, the community of German-language libraries is defining how publications should be described with greater precision than by means of natural language so that others can definitively refer to them. This is where the *Integrated Authority File (GND)* comes into play. Harari refers to the nature of humankind as a whole, and how this differs from the character of chimpanzees, for example. The work of cataloguing, the definition of media based on the rules of descriptive and content cataloguing, is far removed from the challenges faced by Homo Sapiens during the Stone Age. And yet, in a sense, it is simply a different section of the same light beam. As a result of the “cognitive revolution”³ that occurred back then, today, we are facing the challenges of the digital transformation. And this too we will master precisely because of our ability to collaborate. This is what we do. In the course of opening the GND to include communities beyond library institutions, one thing has become ever clearer: the GND is much more than just a collection of nine million authority data records on people, places, corporations, conferences, works and subject headings.⁴ It also describes an organisational structure that reflects the state of its current users. It refers back to a certain data model that is based around the needs of its users. It is subject to specific rules and can be regarded as a specialist tool within an specialised environment defined by the requirements of the library community. Yet the new user groups are organised differently. They have other data models. They catalogue the objects of their interest according to different rules and use a different technical infrastructure. And yet they are still very interested in using the GND authority data. They don’t just wish to use the identifiers in their cataloguing work, but also want to be able to create new GND data records when they see a need to do so. They want to become an active part of the GND community. To this end, we need to work together to consider carefully what we can change, and how much, without damaging the core of the GND. This is because everyone wants to preserve its reliable quality. Our task is once more to organise our collaborative efforts in line with a collective intentionality.

¹ Quote taken from Harari 2020, p. 68.

² Harari 2020.

³ On the concept of the “cognitive revolution”, cf. Harari 2012, pp. 11-100.

⁴ The record type *Conferences* in the GND makes particularly apparent how interwoven the GND is with its users in the world of libraries. That is because this record type describes a specific kind of publisher. See all categories in the GND ontology: <https://d-nb.info/standards/elementset/gnd>

An instrument for broadening participation

Opening the GND is like the concert given by an entire orchestra of stakeholders and activities. One instrument in this orchestra, a starting point for a careful adaptation, is the technical environment in which the GND is rooted. It is not the notion of dispensing with the existing technical infrastructure, but much more the idea of offering a parallel infrastructure, that has drawn our attention to the database software Wikibase⁵. Wikibase is a piece of open-source software from the Wikimedia Foundation. This foundation has previously developed the Mediawiki software, which is used to operate millions of Wikis around the world. The most famous Wiki is Wikipedia, operated by Wikimedia. The Wikidata project was launched nine years ago with the aim of improving Wikipedia. A database for structured data with which one can describe the world in a way that can be read by both humans and machines alike. The software empowering Wikidata is Wikibase. Wikibase features certain properties designed to make large-scale collaboration easier:

- It offers web-based access.
- It facilitates parallel collaborative working.
- It automatically logs the version history and its editors.
- It offers a dedicated discussion page for every data record.
- It is geared towards multilingual user communities.
- It offers a simple and flexible (though also limited) data model.
- Entering new content works easily and intuitively.

We intensively studied these properties at the German National Library in 2019 and summarised our conclusions in our evaluation⁶ in collaboration with Wikimedia Deutschland. In this context, we also explored current weaknesses in the system and potential areas for development. In its current iteration, the system falls far short of meeting all the requirements for an ideal editing system and hub for cultural institutions. To this end, it still needs to outgrow its origins as a piece of Wikidata software. Nevertheless, we were able to identify the fundamental prerequisites for its productive use in the context of the AfS. What matters is less the current status of the product and more the inherent potential in its further development and the establishment of a broad community in the cultural sector.

In 2020, we first considered how to make the most effective use of Wikibase in broadening participation in the GND, before creating the conditions for implementing our plans as efficiently as possible. We decided to become active on three levels. We want to:

- Create a second home for the GND as an authority file within a Wikibase database. New user communities can make suggestions for new GND data records more easily and independently of the existing technical structures, and compare their data to the GND with greater ease in order to avoid duplication.
- Create partnerships with Wikimedia and other institutions also wanting to use Wikibase, in order to collaborate on improving the software so as to ultimately establish an ecosystem for cultural data and research data.
- Thirdly, we want to re-order the very frameworks underpinning the GND and our cataloguing work, make these more accessible and easier to adapt to any changes.

⁵ Link to the Wikibase website: <https://wikiba.se/>

⁶ Link to the blog post on our evaluation: <https://wiki.dnb.de/pages/viewpage.action?pageId=167019461>

The second home of the GND

In the world of libraries, the GND has long served as a referencing and rationalisation tool, as did the four authority files that preceded it. It is integrated into certain frameworks and proprietary software structures that are, however, relatively inaccessible to users from outside the world of libraries. We believe that we can use Wikibase to make it easier for some of these target groups to collaborate on the GND.

To this end, we wish to import all the existing GND data records and their corresponding links into a Wikibase entity in 2021. This may sound like a simple task. However, Wikibase's importation interfaces are still very much aligned with the needs of Wikidata. For this reason, we have sought professional support from a Wikibase specialist, who is assisting us as a service provider in the transfer of the database infrastructure, the data importation and the creation of user-friendly input screens. In the next step, we will then invite experienced and new GND users to test the data-entry and search processes in the new environment so that we can further improve these.

During the second half of 2021, we are planning a technical workflow for synchronising the GND Wikibase entity with the CBS system⁷. The plan is to enable new and existing users without any WinIBW⁸ access to enter their data as a suggestion in the Wikibase entity.

One long-term goal is to offer a user-friendly and supportive data-recording environment for the GND. The *GND web forms*⁹ represent a first step in this direction, as they are considerably more user-friendly than the data-entry systems used by libraries. The web forms currently can be used to record people and corporations. However, this currently envisaged approach is not flexible enough. In addition to the two aforementioned GND record types, there are four more. These six record types unite approximately 50 entity codes¹⁰, each with specific properties via which the respective entities can be recorded as GND data records. These would require a dynamic entry form that adapts to the entity type or usage context chosen, offers necessary and typical entry elements, highlights useful entries and thus guides the user through the entry process. It remains to be seen whether Wikibase represents the right platform for this in the medium term. At present, Wikibase lacks such features. For now, no update to the generic entry interface is planned "ex works". There is also no option of limiting the offering to fundamental elements or values. The user is always confronted with the full range of properties and values, and isn't offered any assistance in decision-making. One aim for 2021 is to establish whether this can be facilitated via the development of a Wikibase expansion, and also which changes would have to be implemented in Wikibase by Wikimedia in order to more adequately support the creation of customisable entry forms that assist the user.

⁷ CBS: proprietary library data-entry software from OCLC.

⁸ WinIBW: licensed software for entering data in the GND.

⁹ The GND web form for persons and corporate bodies is specifically intended for users from cultural institutions such as smaller libraries, archives and museums who would like to create or modify small quantities of data records in the GND. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_Webformular/gnd_webformular.html

¹⁰ Details on entity coding in the GND <https://wiki.dnb.de/download/attachments/90411323/entitaetenCodes.pdf>

The WikiLibrary Manifesto

Another area our work will focus on in 2021 is our partnership with Wikimedia Deutschland and other institutions in order to improve Wikibase as a technical infrastructure. Adherence to the FAIR Data Principles (Findability, Accessibility, Interoperability and Reusability)¹¹ when providing data is becoming increasingly important in an ever-growing number of contexts. Data are to become more interlinked in order to make it easier overall to generate new knowledge. This especially applies to data that were generated using public funding. This represents a great challenge for many institutions. It raises the question as to whether they should offer their data in a collective pool for structured data, like data portals. Such institutions must ask themselves whether they are willing to face all the potential consequences, such as sacrificing control over the data model, data-recording rules and quality-assurance processes. Or should they instead use stand-alone solutions and thus accept that their data will be less visible and get re-used less? By broadening participation in the GND, we wish to create alternatives. We are committed to creating a reliable, machine-readable and communally managed Linked Open Data Network for the arts, sciences and culture as a viable basis for FAIR knowledge. Instead of a central platform, we favour an open network of interlinked databases. This requires a communal organisational framework. We wish to provide this within a single network. A network is only ever as good as the partners within it. To this end, the German National Library co-published the WikiLibrary Manifesto together with Wikimedia Germany. Almost forty institutions have already accepted our invitation. The manifesto invites the undersigning institutions to collaborate on the basis of the following principles:

- Promoting free licenses for data and their software environment.
- Shaping spaces where diverse communities thrive. (Community gardening).
- Providing structured data based on FAIR data principles in order to be able to transparently transform data into information to create FAIR knowledge.
- Promoting common core standards created consensually and collaboratively.
- Providing open governance structures and embedding them into existing systems.
- Dedicating resources to obtain user interfaces that are accessible to and user-friendly for everybody who wants to contribute and actively care for data and knowledge.
- Fostering data literacy in the digital transformation on the three stages: data, information and knowledge.

Of equal importance, if not more so, is the communal implementation of specific measures by all signatories in partnership with Wikimedia Germany. The aim is to promote Wikibase as a promising technical infrastructure for the storage, editing and exchange of data on the basis of the FAIR Data Principles. We wish to shape Wikibase into a user-friendly reference-database software for data hubs in order to promote the desired data ecosystem. To this end, we are inviting further institutions from the world of libraries, from all GLAM (galleries, libraries, archives and museums) areas and from the humanities to use Wikibase in order to create an ecosystem of structured data that comes closer to a true semantic web for FAIR knowledge.¹²

¹¹ Information on the Fair Data Principles https://www.forschungsdaten.org/index.php/FAIR_data_principles

¹² As an institution, you can co-sign the manifesto via a simple form by following this link: <https://www.wikimedia.de/projects/wikilibrary-manifest/>

The DACH documentation platform¹³



Would it have occurred to you that the article opposite about a football match was written by a computer program? In recent years, the results of computer linguistics have evolved ever further with the aid of artificial intelligence. Writing programs draw content from structured databases and construct the texts with the individual components according to certain specifications. This is the backdrop to our deliberations on recording the frameworks for descriptive and content cataloguing¹⁴ and the data-entry guidelines for the GND as structured data within a Wikibase entity. For decades now, we have been issuing extensive, detailed texts with precise instructions on which data fields must be entered in the GND, for example. Underpinning these texts are the frameworks for descriptive and content cataloguing, the requirements and limitations of the respective software used for cataloguing, and ultimately also the requirements for the exchange of data. Each time a detail is amended at any point in this complex network, the same amendment must also be implemented in many texts that refer to said point. In each instance, this requires labour- and time-intensive research in a large number of PDF pages. Another consequence of this form of knowledge management is that lots of detailed information – such as how to enter a date, for example, or how to record a job description, or which code to use for which country – has to be repeated in various places to avoid having to hunt for said information. When making an amendment, it is important to maintain an overview of every other area impacted by that amendment. There is an inherent risk of errors and a lack of transparency. It certainly makes any guidance less user-friendly, as there is a continuous need for amendments.

The basic principle is strikingly simple. Let us first focus on the GND itself. The number of fields with which one can describe entities for authority data records in the Pica or Marc 21 data formats is manageable at around 300. These data fields or elements serve to make statements regarding

¹³ DACH documentation platform: The platform is designed to bring together all the frameworks for library-based cataloguing and the data-entry guidelines for the GND in the German-speaking regions (Germany, Austria and Switzerland).

¹⁴ This refers to the RDA and RSWK frameworks

the properties, relationship types, sub-categories or entity codes for the respective entities being described. The data elements contain defined characteristics and different codes, depending on the data format. If all the elements are stored in a corresponding database, the data elements can be assembled in modular fashion, just like a construction kit, according to the rules of the frameworks the database is organised around.

The data-entry guidelines for people alone, with all the requisite entity codes in the GND, encompass 46 pages.¹⁵ Yet the elements to be recorded are few. Along with the name, the other primary elements are the date of birth and death, and any links to other databases, such as place names in the form of the place of birth, the place(s) where the individual worked, or similar. For each of the entity codes in the record-type “persons, a new description is provided each time of how the element “place” must be modelled, for example. If these definitions were to be stored in a database, as a rule, one could simply enter the respective element. This means that if the rule governing the characteristics for recording a regional corporation changes,¹⁶ one can change this centrally in a single location, and all other locations where this element is used are automatically updated too. It is the same principle as applied in the authority data records of library catalogues.

We have started to describe in a structured format all the elements used in the GND. To do so, we are adopting the specifications contained in the frameworks. Now the challenge will consist in writing comprehensible continuous texts in which one can sensibly embed the elements. These can then be updated more concisely than before, and also potentially serve as the foundation for the creation of entry forms for the database with all GND data records.

Sometimes it is beneficial to reflect on the sense and purpose of one’s work in order to remain motivated, stay focused and convey to others why this work is important and deserves funding. With this workshop report, we wish to provide you with an insight into our work and the ideas behind it. An exciting time of pioneering work lies ahead of us. This work is made even more interesting thanks to the other, concurrent Wikibase projects in the newly formed consortia of the National Research Data Infrastructure Initiative (NFDI) and other major universal and national libraries in Europe and America, with whom we are in close contact. We will keep you up to date on the latest developments.

¹⁵ Also cf. <https://wiki.dnb.de/pages/viewpage.action?pageId=90411361&preview=/90411361/94831186/EH-P-01.pdf>

¹⁶ A local corporation is an entity code from the group of geographic entities or places.

References

Harari, Yuval Noah. 2013. *A brief history of humankind*. Munich: Dt. Verlags-Anstalt.

Harari, Yuval Noah. 2020. *Sapiens. The birth of humankind. Graphic novel*. Munich: C.H. Beck.

Wikidata: a new perspective towards universal bibliographic control*

Carlo Bianchini^(a), Lucia Sardo^(b)

a) Università degli studi di Pavia, <http://orcid.org/0000-0002-6635-6371>

b) Università di Bologna. Campus di Ravenna, <http://orcid.org/0000-0001-6480-759X>

Contact: Carlo Bianchini, carlo.bianchini@unipv.it; Lucia Sardo, lucia.sardo@unibo.it

Received: 10 April 2021; **Accepted:** 22 May 2021; **First Published:** 15 January 2022

ABSTRACT

Traditional UBC provides for the standardization of bibliographic records, the creation of guidelines dedicated to national bibliographic agencies, the creation of the UNIMARC format, and the curation of authority data. Bibliographic Control has deeply evolved since IFLA theorization during the Seventies of the XX Century, due to the availability of a very large range of new bibliographic tools. At the beginning of the XXI century, UBC is quite different and involves new actors. Among these, Wikidata has a background greatly different from that of libraries as institutions: it is not devoted to bibliographic data, nor it is limited to personal authority control, but its value in AC tools like VIAF and National Libraries authority files is undiscussed. After a presentation on how Wikidata items describe and identify bibliographic entities, the authors underline how the existence, use and reuse of Wikidata affect the way the professional community thinks about UBC. Wikidata is a clear example of the need for a new approach to identification and description, that are deeply intertwined. Secondly, from a Wikidata perspective, the relevance of globally preferred and variant access points is lessened. Moreover, descriptions in Wikidata – although conceptually very similar to the traditional one – present differences and potentialities that a traditional description does not have and cannot have. Also from a theoretical perspective, Wikidata offers a pragmatic way to think globally and act locally. In fact, it shows that there is no need for one standardization of practices for establishing the headings and structure of authority records in one international form; instead, users' convenience can be achieved by a technological infrastructure capable to present to each user the information about an entity in its own language and script. Additionally, Wikidata is the most evident example of the distributed and diffused approach of the semantic web to the issue of the universal identification of the entities. Also, Wikidata identification and description show that authority and bibliographic control must be tackled as just a part of the more general topic of the creation of a knowledge graph of all human knowledge by means of linked open data. Lastly, this objective cannot be achieved only by contribution, cooperation, and networking of large national agencies (as in VIAF), as a larger number of stakeholders must be involved to achieve a UBC also including the full indexing of any kind of scientific communication

KEYWORDS

UBC; Universal Bibliographic Control; Wikidata; Metadata; Description; Identification.

* The authors cooperated in the redaction and revision of the article. Nevertheless, each author mainly authored specific parts of the article: Carlo Bianchini: sections 1.2, 2, 4, and 5; Lucia Sardo: sections 1, 3, 4.2, and 5.

© 2022, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.

1. Introduction: From UBC to the semantic web

The idea of bibliographic control, that is, the idea of being able to have an exhaustive overview of what is published (“the mastery over written and published records which is provided by and for the purposes of bibliography”; Unesco and LC Bibliographical Survey 1950, 1), at least as old as bibliography, took on new connotations in the 1970s, with the birth of IFLA’s Universal Bibliographic Control program.

The main objective of universal bibliographic control is the availability of bibliographic records of publications produced in all countries. In the context of the program for the UBC, the emphasis is placed not only on the universality of such control, but also on the standardization of the content of bibliographic records, on the need to have a specific program dedicated to this objective to foster cooperation at the international level and to achieve the goal of a worldwide record of what is published, and on the importance given to the rapid availability of these data.

The program for the UBC has its basis in standardization and in the direct involvement of national bibliographic agencies, a fundamental element for international cooperation: during the 1977 congress devoted to national bibliographies, their tasks were defined to be the documentation of national editorial production, and the drafting of authority records for national authors.

The principles on which universal bibliographic control is based, which have been progressively brought into focus through studies and initiatives, can be summarized as follows: first, the aims of the system are the control and exchange of bibliographic information; worldwide coverage is guaranteed by the cooperation of the national components of the system; the main objective is to make the bibliographic data of all publications universally and promptly available, in an internationally accepted standard format. To achieve this goal, the complete bibliographic record of each publication should be made once only in the country of origin by a national bibliographic agency, in accordance with international standards that permit exchange. In this perspective, the national bibliographic agency, usually established in national libraries, which usually benefit from the mandatory deposit of printed matter, results to be the most appropriate structure for authoritatively identifying and recording the authors and publications of each country, and is responsible for producing a current national bibliography, in which to publish such records as soon as possible, and for distributing such records in various standard formats. The agencies then come to be integrated into an international system and regularly exchange records made.

For UBC principles to be implemented and scaled up, some requirements must be satisfied:

- a canon of principles, standards, and practices governing the creation and structure of catalographic data that is shared on a broad scale must be available.
- each national bibliographic agency must fulfil its responsibilities in a manner that is inclusive of and consistent with accepted standards.
- an infrastructure is needed to support the efficient exchange of data among national bibliographic agencies.

While IFLA’s work on satisfactorily scaling up the UBC concept regarding bibliographic records has been successful, regarding authorities it has been largely driven by the recognition of the need to deal with these three critical factors:

- the standardization of practices for establishing the headings and structure of authorities.

- the promotion of national responsibilities for the creation and “dissemination” of authority records.
- the planning of an infrastructure that supports the effective international exchange of authority records.

The results of the program’s work for the UBC are there for all to see: the publication of ISBD, the creation of the UNIMARC format, the publication of authority lists and tools for controlling the forms of personal and collective names, and the Guidelines for the National Bibliographic Agencies and the National Bibliography. Basically, all the work concerning the standardization of bibliographic descriptions and authority records had its basis in the concept of Universal Bibliographic Control.

1.1 UBC for bibliographic and authority data

From the point of view of the UBC, the ISBD standard was first developed, which had the double function of establishing which data were relevant for bibliographic description and in which order they should be presented. This was the first time that such a standardization effort was undertaken; it preceded the formalization of the program but provided for it as a *conditio sine qua non* for its dissemination.¹

In the same period MARC, a machine-readable format, was created for the exchange of cataloguing information. IFLA, too, considered it essential to develop an international MARC format capable of supporting the exchange of bibliographic data, which is why the development of the UNIMARC format was undertaken, both for bibliographic and authority data. All of this was born, we recall, in a context where catalogs were paper-based, and national needs trumped those of internationalization, especially about the choice of name form for access points.

As Gorman summarizes, “In sum, arriving at a standard set of elements in a standard order and delimited in a standard manner was in the mutual interest of the effort to achieve an international standard for bibliographic description (what became the ISBD); MARC; [sic] and the use of both, each in accord with the other, in achieving national and international standardization, cooperation, and sharing; leading, ultimately, to Universal Bibliographic Control” (Gorman 2014, 826-827).

1.2 Wikidata: a tool of bibliographic interest in the semantic web

In 2011, the Library Linked Data Incubator Group, a working group with the aim “to help increase global interoperability of library data on the Web”,² published its final report. It was focused on what libraries can do for the semantic web and what the semantic web can do for libraries, and it underlined that libraries had created and curated a relevant amount of rich data that can “help reduce redundancy of bibliographic descriptions on the Web by clearly identifying key entities that are shared across Linked Data” (W3C Incubator Group 2011). The report offered a new perspective on thinking about the relevance, scope, and purpose of Universal

¹ For an overview about the origins of ISBD, see (Anderson 1974; Gorman 2014).

² <https://www.w3.org/2005/Incubator/lld/>

Bibliographic Control (UBC), beside to “make universally and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications in all countries” (Anderson 1974, 11).

Since the publication of the Report, many tools with a top-down approach have been developed for the identification of entities (people, locations, works, and expressions) such as VIAF, ISNI or ORCID. The top-down approach of these tools reflects the role assigned to the national agencies by UBC. Nevertheless, some authors suggest that, in the semantic web, “building collaborative authority registries linked to standardised identifiers is one of the fundamental cornerstones of the new Universal Bibliographic Control” (Illien and Bourdon 2014, 15) and that “a better mix of bottom-up and top-down methodologies” is needed to support all those who wish to think globally and act locally (Dunsire and Willer 2014, 11).

Since 2012, Wikidata has developed as a new global actor of the semantic web with both a bottom-up and very inclusive approach. Wikidata is a freely available hosted platform that anyone – including libraries – can use to create, publish, and reuse LOD (Allison-Cassin and Scott 2018). Its main goal and function are to work as a central storage for many Wikimedia projects, but it is also used in external services, for example in VIAF or in the Google Knowledge graph (Vrandečić and Krötzsch 2014), for the enrichment of the quality of bibliographic records (Nguyen, Dinneen, and Luczak-Roesch 2020), and for bibliometrics projects and tools (Lemus-Rojas and Odell 2018; Nielsen, Mietchen, and Willighagen 2017; Hernández-Cazorla, Ramírez-Sánchez, and Rodríguez-Herrera 2019; Seidlmayer et al. 2020; Mietchen and Rasberry 2020). Moreover, in the last years, the Wikidata role as an important tool for identifying entities has been increasingly reconsidered (Association of Research Libraries 2019, 27; van Veen 2019; Linked Data for Production 2020).³

2. Identification in Wikidata

As Wikidata is a central storage for all Wikimedia projects, it aims to record data about any kind of item (i.e., entity) and property relevant for all its projects. For example, items of Wikidata can be geographical places, administrative units, events, architectonic objects, any entity of interest for the user, and, of course, any ‘res’ provided for by IFLA LRM model.

In fact, Wikidata shows a relevant interest for the bibliographic universe. Statistics show that Wikidata records about 91 million of items, 31,5% (ca 22,5 million) of which are scholar articles, and nearly 9% (ca 6.376.000) of the existing items are of human type (Q5). Anyway, this class includes any kind of humans, such as kings, politicians, football players and so on, and not just authors of literary or scientific works. Nevertheless, items representing an authority record (not just of humans) can be estimated to be around 6,3 million.⁴

For identification purposes, Wikidata assigns to each item both an URI – for example, <https://www.wikidata.org/wiki/Q12418> – and a label, a description, and one or more aliases (figure 1).

³ https://www.wikidata.org/wiki/Wikidata:Wikidata_for_authority_control

⁴ Personal items with a VIAF identifier are about 2,3 million, but the number of personal items containing at least one identifier of any VIAF source (such as ISNI, LC, GND etc.) are about 6,3 million.

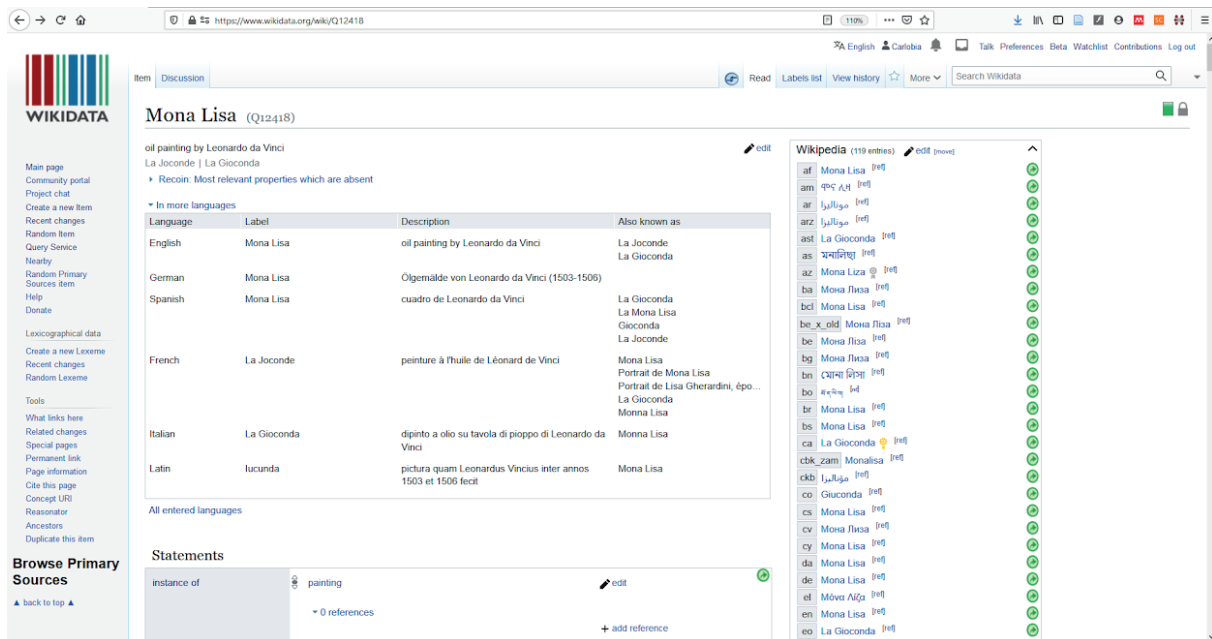


Fig. 1. Example of the main identifying parts (label, description, aliases) of a Wikidata item

The label is the first data element, and it can be considered as the preferred form of the name for the represented entity. In fact, it can be expressed in any existing language and any registered user is enabled to visualize the label in his/her own language and script, if available. Moreover, the preferred form of the name is expressed directly by the user that usually creates the item. So, preferred forms in Wikidata are *literally* founded on common usage and on the convenience of the user provided for by ICP (IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code 2016), and not on these principles *interpreted by a code* of national or international rules! It must be noted that multiple languages and scripts are available for the very same entity, and not for a cluster of nationally created forms like in VIAF.

The second element is the description of the item, that is a short phrase to describe the item. It is in free language and it is useful to quickly distinguish an item from any other item with the same label (for example, “Love”; figure 2), i.e., to disambiguate homonyms.

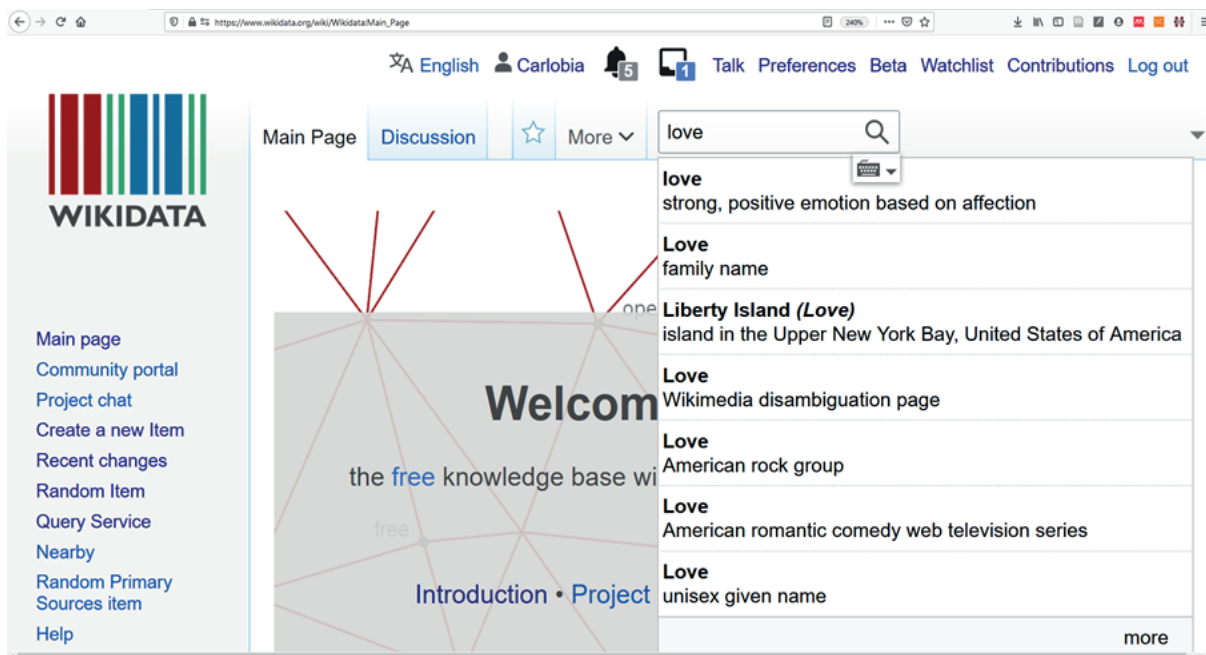


Fig. 2. Descriptions helps to disambiguate items with the same label “Love”

The third element for the quick item identification in Wikidata are the aliases, that are variant forms of the name in one specific language and script (as variant forms of the name in any other language are provided in the form of preferred and variant names in those languages; figure 3).

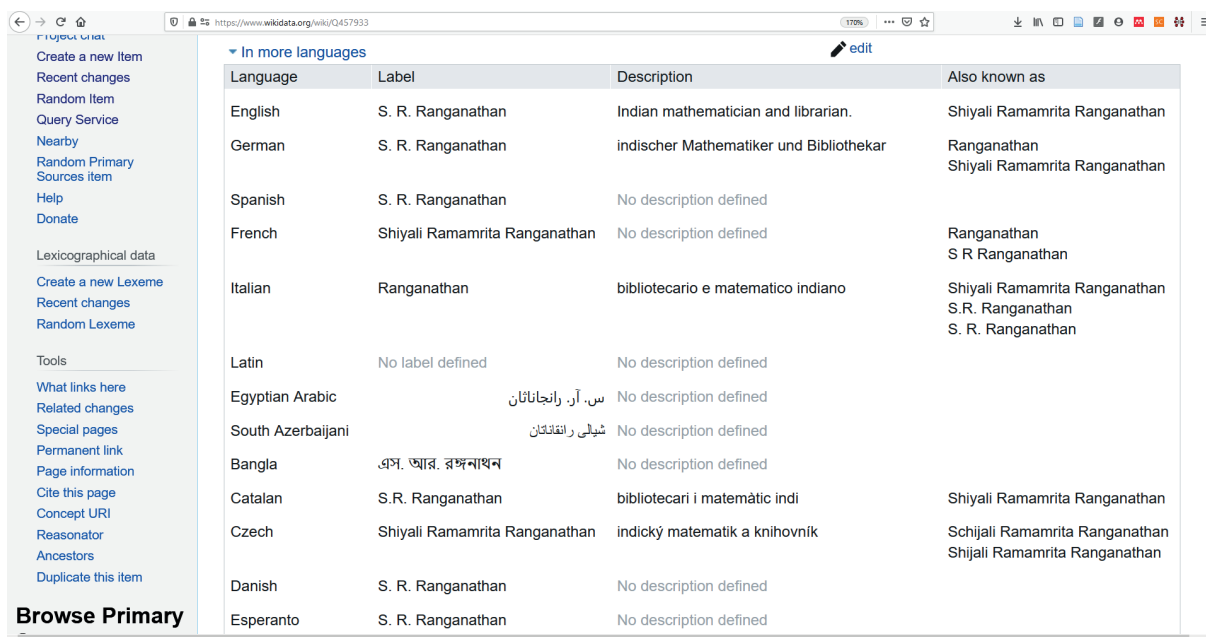


Fig. 3. Aliases available in multiple languages and scripts for S.R. Ranganathan

While the unique identification of the entity is based on a neutral URI (for example: <https://www.wikidata.org/wiki/Q1334284>), both labels and aliases – in any available language and script – work as access points. This pragmatical approach overcomes the theoretical ICP and RDA distinction between preferred and variant access points.

All the remaining properties are registered after the identification elements described above. Nevertheless, they are logically divided into main parts: properties and identifiers. While *properties* are traditionally associated with the *descriptive* goal of the data (see below § no. 3), all the other external identifiers respond to the need of the fourth linked data principle stated by Tim Berners-Lee: “Include links to other URIs so that [users] can discover more things” (Berners-Lee 2006).

External identifiers have the goal to interlink the URI of the item of Wikidata with any other identifiable entity described in the semantic web.⁵ For this reason, Wikidata is more and more recognised for its relevance in the identification of semantic web entities (Association of Research Libraries, 2020; Linked Data for Production, 2020; van Veen 2020). Enriching data with Wikidata ids allows to discover other sources of data and information available in the semantic web.

To create a link towards an external identifier, a specific property must be created in Wikidata to define that identifier. So, it is possible to know how many identifiers are available for different kinds of entities. In figure 4, created by Simon Cobb (Cobb 2019, 5), the number of identifiers associated with each kind of entity are shown.

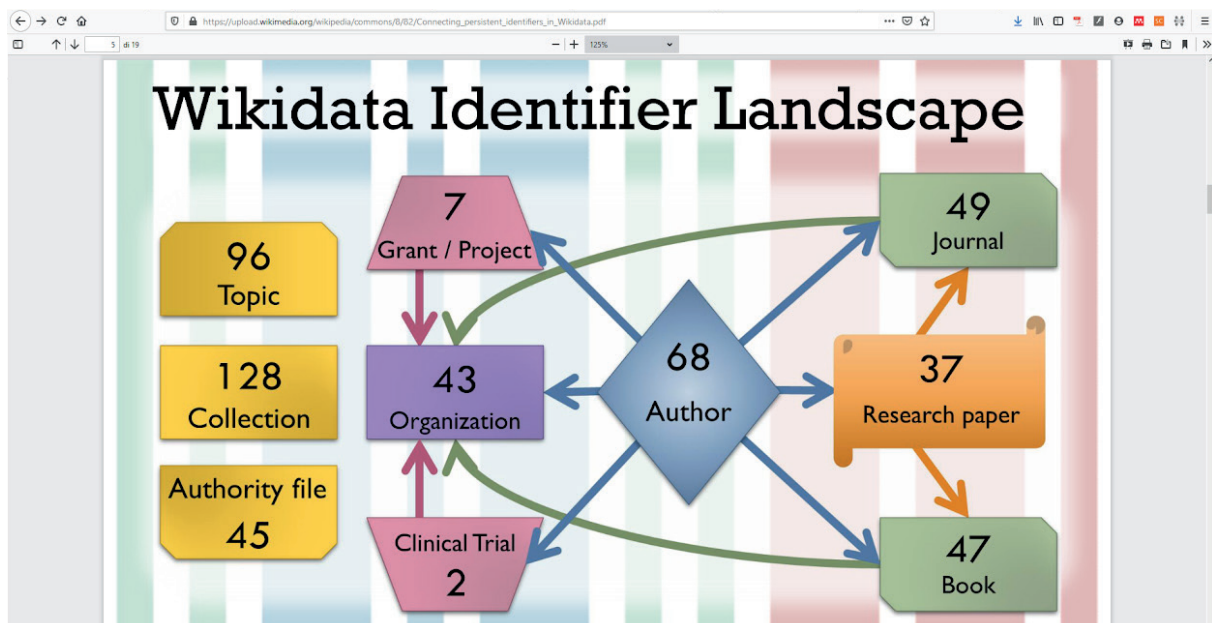


Fig. 4. Number of identifiers in Wikidata for each kind of entity (by Cobb 2019)

⁵ Entity Explosion is a very interesting tool to understand the potential uses of this WD function for the navigation in our discovery tools; see <https://chrome.google.com/webstore/detail/entity-explosion/bbcffeclligkmfocanodamdjclgejcn>.

Most frequently used identifiers in Wikidata are: PubMedID (60,152,490), DOI (26,816,446), PMCID (11,339,676), SIMBAD (8,159,240) and VIAF (6,050,830).⁶

Actually, a major role of Wikidata as a hub for identification in the semantic web is recognised by the VIAF. In fact, VIAF uses Wikidata as an ‘other data provider’, i.e., a provider of data other than a National bibliographic agency. Among the Wikidata items with a VIAF identifier, most common identifiers registered in the personal items are, in decrescent order: ISNI (1,136,260; 18%); DBN (1,012,493; 16%); LC (983,206; 15%); NTA (480,580; 7%); SUDOC (431,919; 6,8%) and BNF (428,792; 6,8%) (Bargioni, Bianchini, and Pellizzari 2021, table 5). The relationship between Wikidata and VIAF is very strong. Wikidata uses property constraints to discover possible inconsistencies in statements both within Wikidata and in the external sources.⁷ So, Wikidata users can check the issues and try to fix them, but any external service can take advantage of this characteristic too.

Identification in Wikidata is a process oriented to the quality of data. First, Wikidata explicitly requires – with the second notability criterion – that each item refers to “a clearly identifiable conceptual or material entity. The entity must be notable, in the sense that it can be described using serious and publicly available references”.⁸ For example, notability prevents Wikidata from accepting isolated clusters formed by VIAF based on a single contributor identifier. Second, clusters of identifiers in a Wikidata item are created by common users and not by automatically performed matches. Matches may be performed semi-automatically (by means of tools such as OpenRefine⁹ or Mix’n’match¹⁰) but human control is always required. Moreover, as in authority work, references are mandatory for each triple and reference sources include encyclopedias, biographical dictionaries, scientific books and articles, in addition to VIAF and other national libraries authority data. More and more Wikidata initiatives are oriented to improve the quality of authors’ data. An example is offered by the bots: a bot in Wikidata “is a program that is allowed to upload large scale data and that is quality controlled by the community” (Siedlmayer 2020). For example, during SWIB 2020 OrcBot was presented: it is a tool created to take advantage of ORCID ids to improve the recording of the property that links the author items to their respective papers, based on ORCID Ids. The Enhancing author items process and issues of reconciliation between ORCID and Wikidata were the focus of the talk “Author items in Wikidata” at the WikiCite Virtual conference 2020 by Simon Cobb – wikimedian in residence at the National Library of Wales.¹¹

3. Description in Wikidata

The “traditional” bibliographical description, marked by descriptive areas in ISBD, and fields and subfields of the MARC format, was firstly challenged by the birth of electronic catalogs, or rather

⁶ https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all, visited 15 December 2020.

⁷ Wikidata helps in identifying issues by two approaches: unique value violations and single value violations. A detailed description of both the approaches and their practical relevance as a quality control tool applied to VIAF is available in (Bargioni, Bianchini, and Pellizzari 2021).

⁸ <https://www.wikidata.org/wiki/Wikidata:Notability>

⁹ <https://openrefine.org/>.

¹⁰ <https://mix-n-match.toolforge.org/>. See also (Agenjo-Bullón and Hernández-Carrascal 2020).

¹¹ https://upload.wikimedia.org/wikipedia/commons/7/79/Author_items_in_Wikidata.pdf

by the new, previously impossible, opportunity to search in any part of the description. Moreover, the electronic catalog challenges the need for a layout made up of descriptive areas, because the flag format of visualization, highlighting the metadata/data structure, overcome the value of the order of citation and the semantics of punctuation.

With the evolution of electronic catalogs and the gradual occurrence of major commercial players (at this stage) into the world of libraries and catalogs, we have therefore come to have tools that allow research and access to different databases, produced by different parties with different purposes. But in this situation the standardization of description is progressively fraying and being lost, in favor of a greater speed in the availability of information about resources, often produced directly by those who produce and make available the resource themselves, whether in analog or digital format.

However, this advantage is detrimental on the search side, as it increases the noise in catalogs and discovery tools, and an insufficiently skilled user may find it difficult to disentangle the results obtained. The impetuous technological evolution of the 21st century, together with a reflection on the functions and object of cataloguing that has led to radical changes in the way we approach resources, is beginning to show the consequences of all this in the cataloguing world. The semantic web and linked data are influencing the ways in which bibliographic data (in the broadest sense) are created, shared, and made available to potential users. In this phase, moreover, no-profit stakeholders outside the world of libraries are becoming increasingly present. An example is Wikidata, created and implemented by a community of volunteers with different training and mindset than those traditionally linked to the book professions, and by volunteers who deal with bibliographic data management.

On the side of libraries, traditional standards are losing their central role in bibliographic description. The static and linear MARC format, subject to many criticisms for many years, is giving way, with difficulty, to different models, such as BIBFRAME, which offers greater flexibility in line with the developments of the semantic web and linked data.

The ISBD format, still used as a standard for describing resources in some cataloging standards, is marking time for the moment. It remains the basis for both the practice and the teaching of cataloging in some settings and situations, but it cannot be denied that we are moving towards other newer and wider ways of describing, like RDA (Resource Description and Access).

RDA, in its first version still linked to AACR2, despite its innovativeness; but, in the new official version from December 2020 has radically changed the way to approach the description of resources. It uses IFLA LRM as a basis for its implementation, with some adaptations that the editors have considered essential for the “practical” needs of the library community, in line with what is expressed in the model, namely that implementations and changes are possible while respecting the basic structure.

On the one hand, RDA allows different levels of description with different types of data encoding (from the mere literal transcription to the use of IRI), on the other hand, it enables libraries using or wanting to use it to adopt different possibilities of use and implementation, with the only constraint to remain “faithful” to the framework and the general choices proposed by the standard and therefore to be interoperable with other realities that use it.

Anyway, an ethical problem must be highlighted: ISBD is a free descriptive standard, while RDA is not; if ISBD disappears, what will remain to those who cannot afford access to RDA?

The description in Wikidata, on the other hand, although conceptually very similar to the traditional one, presents differences and potentialities that a traditional description does not have and cannot have.

As said, conceptually the basic elements of a description are those we are used to in the world of libraries, but we can immediately highlight some important innovations to increase the potential of the description. First, the full implementation of the modelling of the bibliographic universe of IFLA models, with the representation of Works, Expressions, Manifestations, and Items. Secondly, the possibility offered by Wikidata to qualify data. Finally, the possibility of integrating in the description identifiers of different types coming from different sources. While in traditional bibliographic description data qualification was impossible, in Wikidata this is not only feasible, but advisable. In this way you can achieve great advantages for all types of resources that you want to describe. Qualifying is not just specifying the data sources, but their chronological or geographical context; for example, the period of use of a form of a printer's name, a form of a place name, or the language used is a major advantage and a potential that has yet to be fully exploited.

Another aspect relevant to the concept of UBC is the possibility of going in depth in the description of resources. By this we mean that if the UBC very often stopped at the monographic level, in Wikidata instead it is possible to find descriptions of journal articles, or "sheets" of conference proceedings or miscellany. Indeed, perhaps because this lack is significant in traditional catalogues, these types of resources represent a very high percentage of the items in Wikidata.

Certainly, some aspects need to be improved, such as the correct attribution of properties to the right level and the creation of relationships, for example, between works, expressions, events, and items, but the potential is great and the foundations are sufficiently solid to be able to think of continuing the construction of a valid tool for a new vision of UBC, not tied to national conditioning or commercial logic.

The challenges that will have to be faced, at another level, will instead be those related to the use, and the visualization/reuse of these data, but this is not the place to delve into the matter.

The screenshot shows the Wikidata page for the item 'The language of the cataloguer (part 1). The author' (Q58379188). The page is in Italian. The main content area displays a table of language labels and descriptions:

Language	Label	Description	Also known as
English	The language of the cataloguer (part 1). The author	journal article from 'Bibliothecae.it' published in 2017	
Italian	La lingua del catalogatore (parte 1). L'autore	articolo scientifico	
French	No label defined	No description defined	
Sardinian	No label defined	No description defined	

Below the table, there is a 'Statements' section with the following properties:

- instance of**: scholarly article (0 references, + add reference, + add value)
- title**: The language of the cataloguer (part 1). The author (English) (0 references, + add reference, + add value)
- main subject**: cataloging (0 references, + add reference)

On the right side of the page, there is a list of linked Wikidata items:

- Wikipedia (0 entries) [edit]
- Wikibooks (0 entries) [edit]
- Wikinews (0 entries) [edit]
- Wikiquote (0 entries) [edit]
- Wikisource (0 entries) [edit]
- Wikiversity (0 entries) [edit]
- Wikivoyage (0 entries) [edit]
- Wiktionary (0 entries) [edit]
- Multilingual sites (0 entries) [edit]

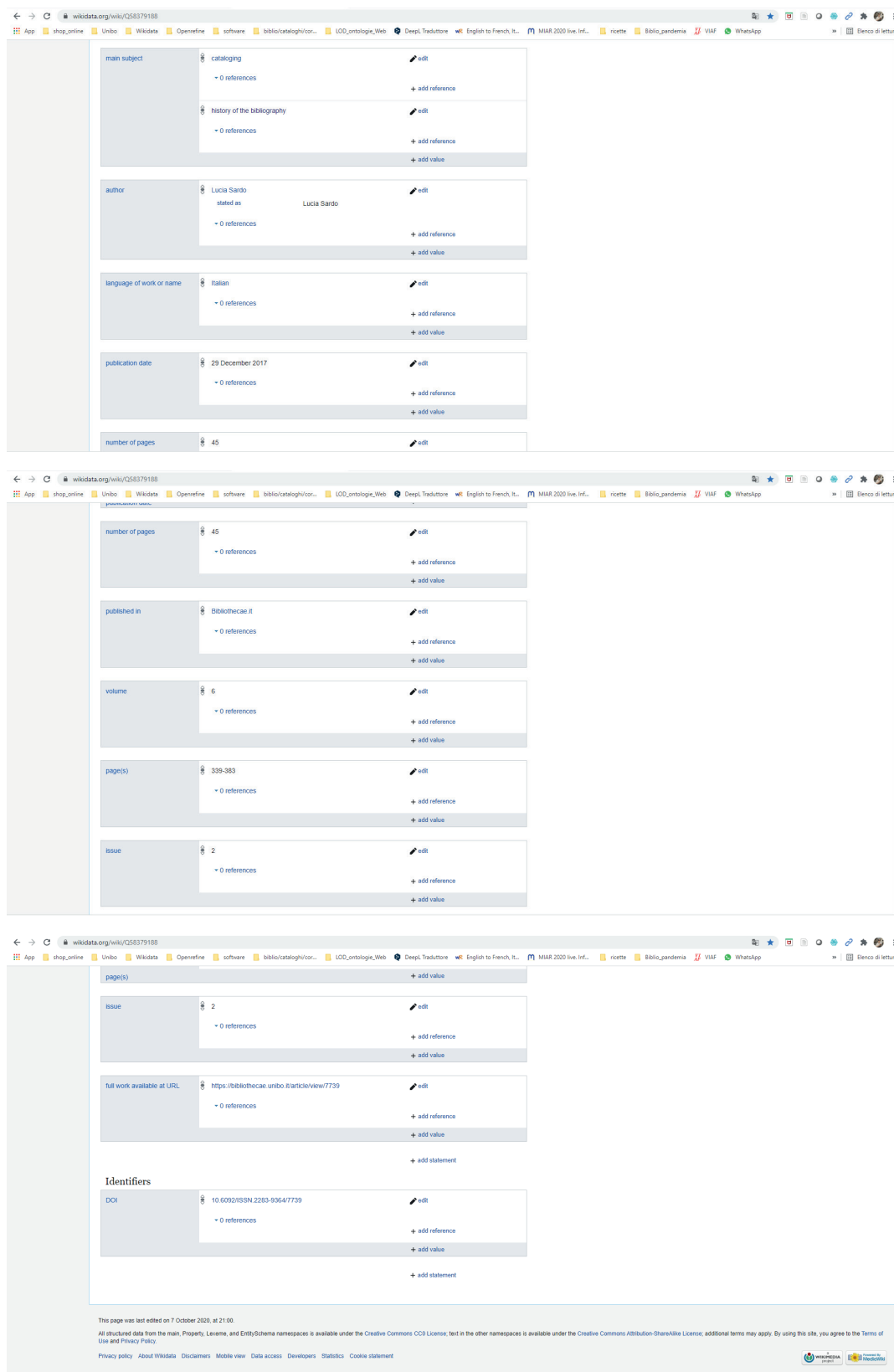


Fig. 5. Example of a Wikidata description of a scholarly article. <https://www.wikidata.org/wiki/Q58379188>

The image shows a browser window displaying the Wikidata template `Template:Bibliographic properties`. The page is organized into several sections, each with a list of Wikidata properties and their corresponding labels in Italian. The sections include:

- Scholarly articles:** Properties like `author` (P50), `copyright license` (P272), `article ID` (P2322), `DOI` (P368), `PMCID` (P302), `arXiv ID` (P1616), `ADS bibcode` (P1616), `Zotero publication ID` (P2027), `JSTOR article ID` (P588), `SSRN article ID` (P693), `NIOSHTIC-ID` (P2388), `Dialnet article ID` (P1610), `CNII article ID` (P2409), `OpenCitations bibliographic resource ID` (P2181), `ACM Digital Library citation ID` (P3332), `IEEE Xplore document ID` (P2405), `Publons Publication ID` (P3411), `Semantic Scholar paper ID` (P4011), `Google Scholar paper ID` (P4028).
- Scholarly journals:** Properties like `title` (P1478), `publisher` (P123), `place of publication` (P291), `editor` (P26), `copyright license` (P272), `language of work or name` (P407), `part of series` (P176), `publication date` (P57), `main subject` (P221), `ISSN` (P229), `SUDOC editors` (P1025), `ZDB ID` (P1942), `JSTOR journal ID` (P1226), `Nawraajee Register journal ID` (P12705), `librisid` (P1226), `Dialnet journal ID` (P1610), `HealthTrust ID` (P1614), `Perseus journal ID` (P2722), `Uniz Review journal ID` (P2725), `Research Papers in Economics Series handle` (P2781), `EBE journal ID` (P2781), `Latindex ID` (P3123), `ERIH PLUS ID` (P343), `EPH journal ID` (P486), `Harvard botanical journal ID` (P474), `Tropicos publication ID` (P484).
- Proceedings series:** Properties like `title` (P1478), `language of work or name` (P407), `editor` (P26), `volume` (P178), `part of the series` (P176), `publication date` (P57), `main subject` (P221), `full work available at URL` (P503).
- Proceedings series:** Properties like `title` (P1478), `language of work or name` (P407), `main subject` (P221), `full work available at URL` (P503).
- Supplements:** Properties like `data` (P28), `editor` (P26), `sponsor` (P39), `published in` (P1433), `supplement to` (P2334).
- Theses:** Properties like `data` (P28), `title` (P1478), `language of work or name` (P407), `main subject` (P221), `full work available at URL` (P503), `dissertation submitted to` (P4191), `thesis committee member` (P1610).
- Books (more):** Properties like `author` (P50), `translator` (P355), `editor` (P26), `illustrator` (P115), `publisher` (P123), `title` (P1478), `subtitle` (P1680), `copyright license` (P272), `full work available at URL` (P503), `language of work or name` (P407), `main subject` (P221), `Commons category` (P373), `copyright holder` (P291).
- Authors (more):** Properties like `name in native language` (P1530), `affiliation` (P1415), `e-mail address` (P268), `field of work` (P101), `notable work` (P300), `Commons category` (P373), `author citation (zoology)` (P335), `ORCID ID` (P405), `VIAF ID` (P14), `ZoteroBank author ID` (P2008), `Scopus author ID` (P1185), `Google Scholar author ID` (P1985), `ResearchGate profile ID` (P2038), `ISNI` (P213), `Dialnet author ID` (P1607), `Perseus author ID` (P2722), `Uniz Review author ID` (P2734), `ACM Digital Library author ID` (P2404), `DBLP author ID` (P2405), `Twitter username` (P2002), `SlideShare username` (P4018), `CNII author ID (books)` (P271), `CNII author ID (articles)` (P478), `GONAT author ID` (P211).
- Publishers:** Properties like `headquarters location` (P130), `field of work` (P101), `Commons category` (P373), `SlideShare username` (P4018).
- Funders:** Properties like `DOI prefix` (P1602), `ISSN publisher prefix` (P2035), `Directory of Czech publishers ID` (P4840).
- Other:** Properties like `describes a project that uses` (P4918), `Chronising America newspaper ID` (P4868), `publication in which this taxon name was established` (P5326), `Archives West finding aid ID` (P4238).

Fig. 6. Wikidata template with all the properties for the bibliographic description – sample; cfr. https://www.wikidata.org/wiki/Template:Bibliographic_properties

Work item properties [edit]					
These properties are for the first item that represents a book (FRBR work level). They are mainly meant to be used for items linked to Wikipedia pages .					
Works should be instances of written work (Q47461344) or one of its subclasses.					
Title	ID	Data type	Description	Examples	Inverse
instance of	P31	Item	instance of: that class of which this subject is a particular example and member	The <i>Autobiography of Alice B. Toklas</i> <instance of> written work	-
title	P1476	Monolingual text	original title and title: published title of a work, such as a newspaper article, a literary work, a website, or a performance work	<i>Diary of Anne Frank</i> <title> Het Achterhuis	-
subtitle	P1680	Monolingual text	subtitle: for works, when the title is followed by a subtitle	<i>Diary of Anne Frank</i> <subtitle> Dagboekbrieven van 12 Juni 1942 – 1 Augustus 1944	-
author	P50	Item	author and writer: main creator(s) of a written work (use on works, not humans), use P2093 when Wikidata item is unknown or does not exist	Harry Potter and the Philosopher's Stone <author> J. K. Rowling	-
possible creator	P1779	Item	creator: for a creative work with considerable uncertainty about the author	Rampin Rider <possible creator> Rampin Master	-
contributor to the creative work or subject	P767	Item	contributor: person or organization that contributed to a subject: co-creator of a creative work or subject	Discworld Noir <contributor to the creative work or subject> Terry Pratchett	contributed to creative work
editor	P98	Item	editor: editor of a compiled work such as a book or a periodical (newspaper or an academic journal)	Oesterreichisches Biographisches Lexikon 1815–1950 <editor> Austrian Academy of Sciences	-
language of work or name	P407	Item	language: language associated with this creative work (such as books, shows, songs, or websites) or a name (for persons use "native language" (P103) and "languages spoken, written or signed" (P1412))	<i>Diary of Anne Frank</i> <language of work or name> Dutch	-
inception	P571	Point in time	date of establishment: date or point in time when the subject came into existence as defined	Tirant lo Blanc <inception> 1490	-
movement	P135	Item	cultural movement: literary, artistic, scientific or philosophical movement or scene associated with this person or work	Urbi et orbi <movement> Russian symbolism	-
has edition or translation	P747	Item	version, edition, or translation, translated work and source text: link to an edition of this item	Brockhaus Enzyklopädie <has edition or translation> Brockhaus Enzyklopädie (21 ed.)	edition or translation of
form of creative work	P7937	Item	form and album type: structure of a creative work	The Hours <form of creative work> novel	-
genre	P136	Item	genre: creative work's genre or an artist's field of work (P101). Use main subject (P921) to relate creative works to their topic	The Hours <genre> feminist novel	-
main subject	P921	Item	topic, matter and topic of a work: primary topic of a work (see also P180: depicts)	The Party Journalist <main subject> radio propaganda	-
award received	P166	Item	award, honorary citizenship, title of honor and statue: award or recognition received by a person, organisation or creative work	The Hours <award received> Pulitzer Prize for Fiction	-
follows	P155	Item	follows: immediately prior item in a series of which the subject is a part (if the subject has replaced the preceding item, e.g. political offices, use "replaces" (P1385))	A Feast for Crows <follows> A Storm of Swords	followed by

Fig. 7. Wikidata template with the Work item properties; https://www.wikidata.org/wiki/Wikidata:WikiProject_Books#Work_item_properties

4. Wikidata as a bibliographic tool. Strategies, projects, and tools

4.1 Identification

To understand how the identification process can be improved in Wikidata, it is necessary to distinguish two possible approaches: by identifiers and by properties of the items.

The most important tool of quality control for proper identification of items are identifiers.¹² In fact, every property associated with an external identifier is provided with constraint rules – usually, an external identifier must be associated with only one item and one item must have only one identifier per type. These rules are extremely important for bibliographic control. In fact, they allow to identify possible errors within Wikidata (e.g., a duplicated item), but above all, they show possible mistakes also within the sources of the external identifiers linked to a Wikidata item (e.g., when a duplication of external identifiers occurs in a Wikidata item).

An example of how it works can be useful to understand how much Wikidata can help in the identification work for persons. A quick check of the identifiers associated with “Ferruccio Battolini” shows that three distinct VIAF IDs and two distinct ISNI IDs are associated with the same person (figure 8a).¹³ This example is relatively simple, but things get more complicated with classical authors (e.g., poetess Saffo; figure 8b).¹⁴

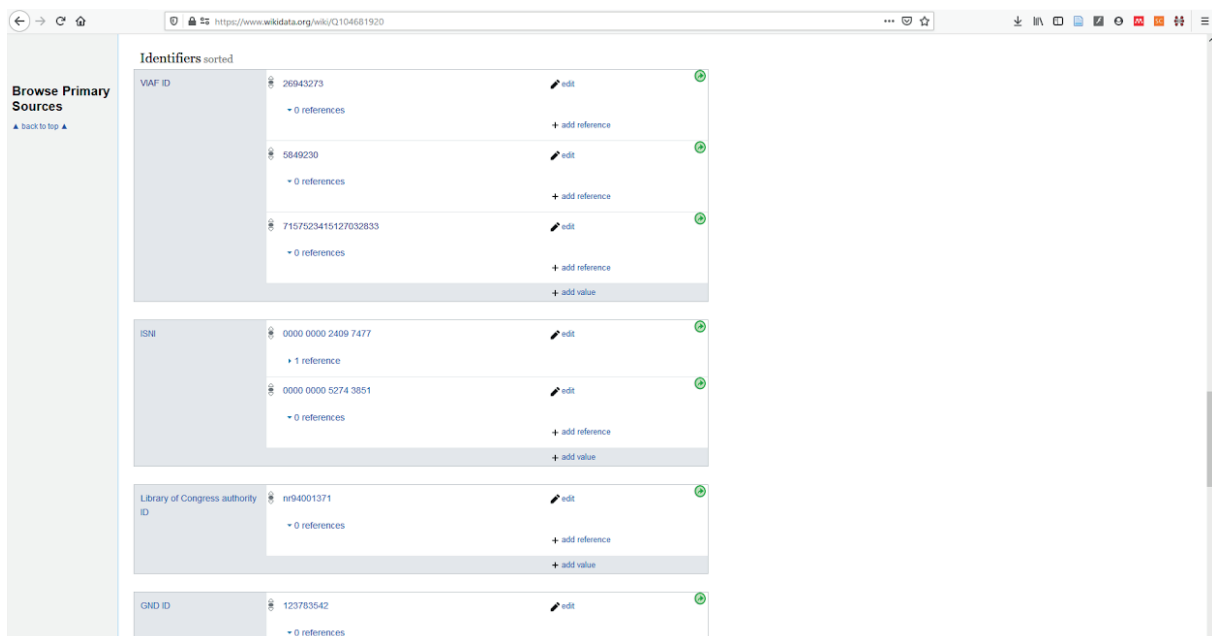


Fig. 8a. Duplicated VIAF and ISNI identifiers for Ferruccio Battolini

¹² See Wikidata Project: https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control.

¹³ <https://www.wikidata.org/wiki/Q104681920>.

¹⁴ <https://www.wikidata.org/wiki/Q17892>.

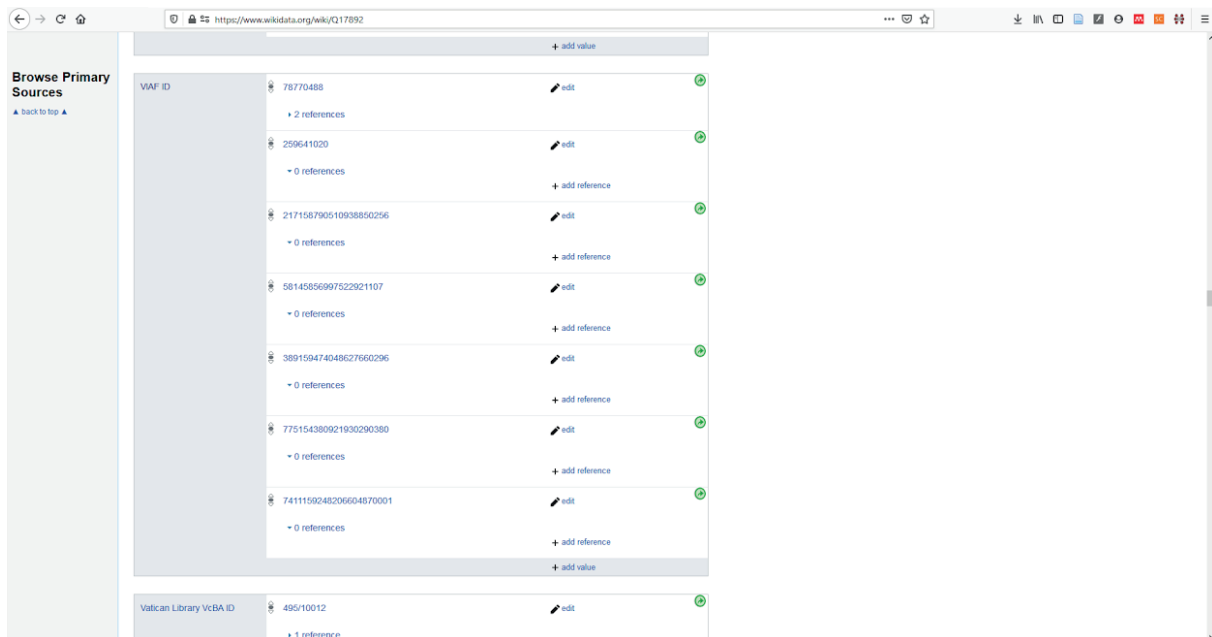


Fig. 8b. Duplicated VIAF identifiers for Sappho

As VIAF remains a major source for Wikidata, the community developed specific tools – named gadgets – to improve the reuse of its data in Wikidata items. Gadgets are enhancements of the edit interface for registered users and are very useful for data production.¹⁵ For instance, the gadget *MoreIdentifiers* was created by Stefano Bargioni and Camillo Pellizzari to facilitate the creation of links between Wikidata items and VIAF entities and it enables users to add easily and quickly authority control IDs from VIAF with few edits checking the identifier and clicking on the button (figure 9).¹⁶ Moreover, it enables to know whether an identifier is old or wrong (as it is presented strikethrough in red) and to create a report for any wrong identifier wrongly included in the VIAF cluster, if the case, by means of the thunder icon. A page of identifiers wrongly included in a VIAF cluster is maintained and constantly updated by Wikidata users; alas, it seems not so used by VIAF managers.¹⁷

¹⁵ A list of gadgets is available at <https://www.wikidata.org/wiki/Wikidata:VIAF/cluster#Gadgets>.

¹⁶ <https://www.wikidata.org/wiki/User:Bargioni/moreIdentifiers>.

¹⁷ https://www.wikidata.org/wiki/Wikidata:VIAF/cluster/conflating_specific_entries.

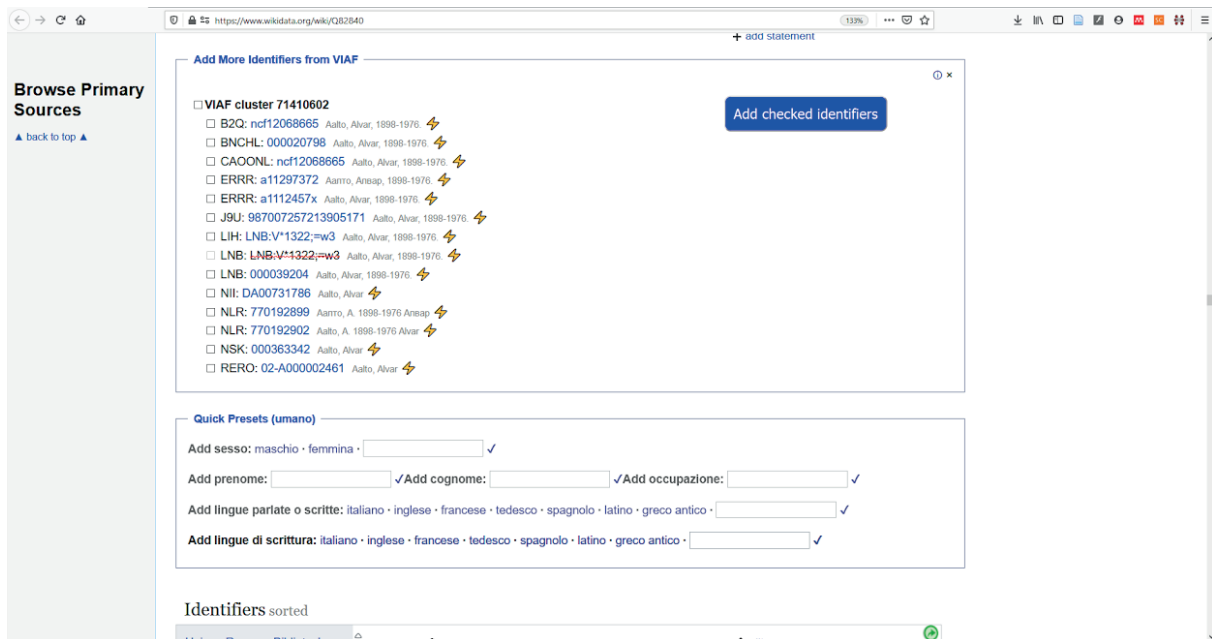


Fig. 9. Box of the Wikidata gadget *MoreIdentifiers*.

MoreIdentifiers works for any kind of VIAF entity (such as geographic names or corporate names) but it is best useful for personal names.

Properties are key for the identification of items within Wikidata. In this case, identification is based on the matching of several properties. For example, human beings' identification is usually based on the matching of the label, the description, and the dates of birth and death.

In this approach, the more the available properties, the more the probabilities for identification. So, the number of properties of an item is a key issue, because a higher number of properties describing an entity assure a more probable identification – or disambiguation – of the two entities being compared.

For this reason, a dedicated gadget was developed by Wikidata community: *Recoin*, i.e., *Relative Completeness in Wikidata* (figure 10). *Recoin* is a “script that extends Wikidata entity pages with information about the *relative completeness* of the information” referring to the “extent of information found on an item in comparison with other similar items”.¹⁸ *Recoin* is a tool to help authors of Wikidata to know on which data attention must be focused on; moreover, it is also extremely useful for data consumers to be aware of the degree of information available about an item.

¹⁸ <https://www.wikidata.org/wiki/Wikidata:Recoin>.



Fig. 10. Example of missing properties highlighted by Recoin

As shown in figure 10, Recoin offers a status indicator icon, ranging from very detailed to very basic, to indicate the relative completeness of the description of an item on a 5-level scale, and a list of the most relevant properties which are not present in the item. Missing properties are detected by a comparison of the properties in that item and the properties most frequently occurring in that class. For example, the properties in an item representing a politician are compared to the most frequently occurring properties of the item belonging to the class ‘politicians’ (Balaraman, Razniewski, and Nutt 2018).

Nevertheless, identification within Wikidata is far from being perfect and can still be improved. Many new items are poorly described because of two main issues: many items are created by semi-automatic processes and, for this reason, data can be incorrect, generic (e.g., string versus author; cf. below),¹⁹ or poor. In addition, at present Wikidata as a semantic web hub, is undoubtedly more oriented towards identifying than describing items.

When data derived from external sources are incorrect, their limit is inherited in the Wikidata item description (as seen above with VIAF identifiers). For example, a large part of the item creation work from sources like ORCID is made by bulk upload from bots; this means that in these cases “errors can persist for many months without being rectified and can be replicated in bulk editing of the description without detection” (Cobb 2020, 3).

External data can result in generic data too. For example, it is possible to import data from Zotero – a Reference Manager Software – to Wikidata, but the authors of the books or the articles are recorded as a *string of characters* (P2093), instead of a relationship between the item and the *author* (P50). And this happens with many other automatic tools, so that about

¹⁹ https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/all.

135 million of authors are recorded as strings compared to just 20 million recorded as author relationships.

Poor external source data can produce a low number of average statements, and this means a poorer description and a more difficult process of identification of the items (mainly towards other external sources). For example, a study by Simon Cobb shows that items having an ORCID have a low number of statements; so that “the latest author items are sparse in comparison to older items, which have had longer to attract the curatorial efforts of community members” (Cobb 2020, 3).

Anyway, as author item relations allow for much richer analysis, to fix the issue, the special tool *Author disambiguator*²⁰ was created: it is a tool for editing the authors of works recorded in Wikidata and for assisting in converting “those strings into links to author items as efficiently and easily as possible”.²¹

Another issue for authority control in Wikidata is that the data harvesting process is not structured; initiatives to upload data are very much, and sometimes based on semi-automatic tools, but there is not a clear overall design nor strategy, as typical of bottom-up approaches.

At the end of his analysis, Simon Cobb suggests a few steps to improve identification process for Wikidata items that can be applied to any kind of item and of external source; major suggestions are:

- Seek community consensus on minimum acceptable standard for author items created by bot imports.
- Define author data requirements for a variety of use cases.
- Review and validate data in existing author items.
- Organise an online workshop to facilitate discussion and collaboration between interested members of the Wikidata editor community and other stakeholders within and outside the Wikimedia Foundation projects.
- Establish a WikiProject special interest group (SIG) to focus on the improvement and maintenance of author items” (Cobb 2020, 10).

4.2 Description

The improvement of description in Wikidata can be approached from at least two points of view: the description of a remarkable variety of items in Wikidata, and the more traditional description of bibliographic resources.

In the first case (item description) the issues are certainly more intertwined with the problems of identification, because to have good descriptions, a correct identification of the entity is necessary and therefore the number and quality of properties necessary and sufficient for the description itself must be defined.

In the second case, instead, the improvement would require the growth of the available resources and their univocal identification when possible (by means of identifiers such as ISBN, ISSN, DOI). Problems arise for all the resources that do not have an identifier univocally assigned by an internationally recognized agency, but that have only identifiers assigned by the world of libraries (BID; etc).

²⁰ <https://author-disambiguator.toolforge.org/>.

²¹ https://www.wikidata.org/wiki/Wikidata:Tools/Author_Disambiguator. See also (Smith 2020)

A key issue is the quality of the source metadata, as digitized resources or “digital libraries” show when collecting the product of digitization from different sources (one example for all, Internet Archive). In such situations, the critical issues concern the description of the “less standardized” events and the need of a proper identification of the expressions and works on the one hand, and on the other, a trustful reconstruction of the physical presentation of the events, to facilitate more specialized or bibliographic research.

The description approach provided by RDA, i.e., based on identifiers and IRIs, is particularly effective in a context like Wikidata, and could lead to a significant growth of the number of described resources, and to an enhancement in the quality of the descriptions, as well as to the correct identification of the various entities.

Compared to this, the advantages related to the possibility of making a more granular and detailed bibliographic control than library catalogs are certainly notable (articles, miscellanea perusal, etc.); the possibility of inserting identifiers related to the catalogs of the major libraries or library systems worldwide also allows to satisfy the user function *to obtain*, which is often what most users want as a result of a search.

Finally, we remember that description and identification issues are inevitably intertwined.

5. Suggestions from Wikidata for the UBC

Wikidata is a Wikimedia tool to meet the needs of Wikimedia platforms, but it has relevant bibliographic features that can help to better understand the future of the Universal Bibliographic Control. In fact, Wikidata offers a completely new approach to data management that involves the way in which our community thinks and operates the Universal Bibliographic Control, both from a practical and theoretical perspective.

Wikidata is not designed as a bibliographic tool, and it is not oriented, nor limited, to bibliographic resources. For this reason, even if a data schema is available as Wikidata property page for works, editions, scientific articles, serials and so on, the quality and completeness of bibliographic data are usually high, but not certain. In fact, the number and the quality of the identifiers and properties recorded in Wikidata items are very varying, and the oldest items are usually more well-structured than the most recent ones; anyway, many gadget and tools (*MoreIdentifiers*, *Recoin*, *Author Disambiguator*, etc.) are available to improve them. Furthermore, while its bottom-up approach is a major asset in a global environment in which the role of great national bibliographic agencies is unable to fulfil the requirements of UBC, it is also a limit for the lack of a clear overall strategy of implementation of authority and bibliographic data.

Anyway, from a practical perspective, Wikidata is a clear example of the need for a new approach to identification and description, that are intertwined. First, a change in the workflow and in the mindset is required to the cataloguer, because a basic and even problematic identification must precede the description of the item.

Secondly, the relevance of globally preferred and variant access points is lessened; in fact, they remain relevant just in a local environment, and in a specific context defined by a particular set of rules. While labels and aliases pragmatically meet the requirements of making data accessible for any users’ search, the identification function – a pillar of UBC – is assured by international

identifiers, among which Wikidata ID is more and more significant. Moreover, the description in Wikidata – although conceptually very similar to the traditional one – presents differences and potentialities that a traditional description does not have and cannot have. First, the full implementation of the modelling of the bibliographic universe of IFLA models is available, with the representation of Works, Expressions, Manifestations, and Items. Third, the possibility of integrating in the description identifiers of different types coming from different sources, above all from the library field. Last, but not least, the possibility offered by Wikidata to qualify data. For instance, the chance to specify the period of use of a form of a printer's name, or of a place name, or of the used language is a major advantage and a potential that has yet to be fully exploited.

Another relevant point in Wikidata practical perspective is its major value as a *de facto* infrastructure to support the efficient exchange of bibliographic data among users, especially those who are not national bibliographic agencies. Wikidata is already a major hub of the semantic web also for bibliographic purposes. Moreover, Wikidata can record and disseminate bibliographic data of analytic descriptions, such as scholarly articles or chapters of books. Finally, Wikidata upgrades the concept of authority work, including reference both to the main international library catalogs and to local library catalogs and to a wider variety of reference sources (such as encyclopedias, dictionaries, and biographical repertoires).

From a theoretical perspective, Wikidata offers a pragmatic way to think globally and act locally. For instance, it suggests looking at authority data as just a part of a wider perspective in which we produce and record data. For instance, it helps to recognize that an 'author' is just a person with a typed relationship toward a work, or a subject is any kind of entity with another typed relationship with a work. Authors and subjects, in a sense, do not exist in 'nature', but they become meaningful only in a bibliographic data perspective, and they must be expressed by a relation of authorship or aboutness between entities.

Furthermore, it shows that there is no need for one standardization of practices for establishing the headings and structure of authority records in one international form; instead, users' convenience can be achieved by a technological infrastructure capable to present to each user the information about an entity in its own language and script. Wikidata is the most evident example of the distributed and diffused approach of the semantic web to the issue of the universal identification of the entities. Moreover, thanks to a bottom-up and co-operative approach, Wikidata fulfils the requirements of International cataloguing Principles of common usage and convenience of the user by means of the users themselves.

There are other two relevant points about the contribution of Wikidata to the theoretical framework of the Universal Bibliographic Control. The first is that authority and bibliographic control must be tackled as just a part of the more general topic of the creation of a knowledge graph of all human knowledge by means of linked open data. In fact, Wikidata and the Semantic Web record data for any kind of item and not just for entities of bibliographic interest. In this new context, data for the achievement of the Universal Bibliographic Control and data, information, resources controlled by the Universal Bibliographic Control are perfectly integrated in one structure. The second is that this objective cannot be achieved only by contribution, cooperation, and networking of large National Agencies, as a larger number of stakeholders must be involved to achieve a UBC also including the full indexing of any kind of scientific communication.

Bibliographic references

- Agenjo-Bullón, Xavier, and Francisca Hernández-Carrascal. 2020. 'Wikipedia, Wikidata y Mix'n'match'. *Anuario ThinkEPI* 14. <https://doi.org/10/gbj6t>.
- Allison-Cassin, Stacy, and Dan Scott. 2018. 'Wikidata: A Platform for Your Library's Linked Open Data'. *Code4Lib Journal*, 4 May 2018. <https://journal.code4lib.org/articles/13424>.
- Anderson, Dorothy. 1974. *Universal Bibliographic Control. A Long Term Policy - A Plan for Action*. Munchen: Verlag Dokumentation.
- Association of Research Libraries. 2019. *ARL White Paper on Wikidata. Opportunities and Recommendations*.
- Balaraman, Vevake, Simon Razniewski, and Werner Nutt. 2018. 'Recoin: Relative Completeness in Wikidata'. In *WWW '18 Companion: The 2018 WebConference Companion*, April 23–27, 2018, Lyon, France. New York, NY, USA: ACM. <https://doi.org/10.1145/3184558.3191641>.
- Bargioni, Stefano, Carlo Bianchini, and Camillo Pellizzari. 2021. 'Beyond VIAF. Wikidata as a Complementary Tool for Authority Control in Libraries'. *Information Technology and Libraries* 40 (2). <https://doi.org/10.6017/ital.v40i2.12959>
- Berners-Lee, Tim. 2006. 'Linked Data - Design Issues'. 27-7-2006. 2006. <http://www.w3.org/DesignIssues/LinkedData.html>.
- Cobb, Simon. 2019. 'Connecting Persistent Identifiers in Wikidata'. In *Portland PID Workshop, 6th May 2019*. https://upload.wikimedia.org/wikipedia/commons/8/82/Connecting_persistent_identifiers_in_Wikidata.pdf.
- . 2020. 'Author items in Wikidata'. Presented at the WikiCiteVirtual Conference, October 26. https://upload.wikimedia.org/wikipedia/commons/c/cc/WikiCite_Virtual_Conference_2020_-_Author_items_in_Wikidata_-_Slides.pdf.
- Dunsire, Gordon, and Mirna Willer. 2014. 'The Local in the Global: Universal Bibliographic Control from the Bottom Up'. In *IFLA WLIC 2014*. Lyon, France. <http://library.ifla.org/817/>.
- Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Esloa, Steven Folsom, et al. 2020. 'Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage'. OCLC. 4 May 2020. <https://doi.org/10.25333/faq3-ax08>.
- Gorman, Michael. 2014. 'The Origins and Making of the ISBD: A Personal History, 1966–1978'. *Cataloging & Classification Quarterly* 52 (8): 821–34. <https://doi.org/10.1080/01639374.2014.929604>.
- Hernández-Cazorla, Iván, Manuel Ramírez-Sánchez, and Gregorio Rodríguez-Herrera. 2019. 'Wikidata, WikiCite y Scholia Como Herramientas Para Un Corpus de Datos Bibliográficos Enlazados. Curación y Estructuración de La Producción Científica de Los Investigadores Del IATEXT'. *PRISMA.COM* 40 (2019): 78–87.
- IFLA Cataloguing Section and IFLA Meeting of Experts on an International Cataloguing Code. 2016. *Statement of International Cataloguing Principles (ICP)*. Den Haag: IFLA.

- Illien, Gildas, and Françoise Bourdon. 2014. 'A la recherche du temps perdu, retour vers le futur: CBU 2.0'. In *IFLA WLIC 2014*. Lyon, France. <http://library.ifla.org/956/>.
- Lemus-Rojas, Mairelys, and Jere D. Odell. 2018. 'Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project Using Wikidata and Scholia'. *Journal of Librarianship and Scholarly Communication* 6. <https://doi.org/10.7710/2162-3309.2272>.
- Linked Data for Production. 2020. 'Wikidata as a hub for identifiers'. Google Docs. 11 June 2020. https://docs.google.com/presentation/d/1jWz3_nCf5rdd-7ejETGlfv99UV2PnD1v/edit?usp=embed_facebook.
- Mietchen, Daniel, and Lane Rasberry. 2020. 'Presenting Scholia. A Scholarly Profiling Tool'. Presented at the LD4 Wikidata Affinity Group, August 11. https://docs.google.com/presentation/d/1jJbYSnYSDh36-LxzSpedFyWUzusZAJuBbP-y46ji-0w/edit#slide=id.g35f391192_00.
- Nguyen, Ba Xuan, Jesse David Dinneen, and Markus Luczak-Roesch. 2020. 'A Novel Method for Resolving and Completing Authors' Country Affiliation Data in Bibliographic Records'. *Journal of Data and Information Science* 5 (3): 97–115. <https://doi.org/10/ghsnkn>.
- Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. 'Scholia, Scientometrics and Wikidata'. In *The Semantic Web: ESWC 2017 Satellite Events*, edited by Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, 10577:237–59. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70407-4_36.
- Seidlmayer, Eva, Jakob Voß, Tetyana Melnychuk, Lukas Galke, Klaus Tochtermann, Carsten Schultz, and Konrad Forstner. 2020. 'ORCID for Wikidata – Data Enrichment for Scientometric Applications'. In *Proceedings of The 1st Wikidata Workshop*. https://wikidataworkshop.github.io/papers/Wikidata_Workshop_2020_paper_9.pdf.
- Smith, Arthur P. 2020. 'Author Disambiguation'. In *WikiCite 2020 Virtual conference*. https://upload.wikimedia.org/wikipedia/commons/3/38/WikiCite_2020_Author_items.webm.
- Unesco/LC Bibliographical Survey. 1950. *Bibliographical Services: Their Present State and Possibilities of Improvement*. Washington: Library of Congress.
- Veen, Theo van. 2019. 'Wikidata: From “an” Identifier to “the” Identifier'. *Information Technology and Libraries (Online)* 38 (2): 72–81. <https://doi.org/10/ghbj62>.
- Vrandečić, Denny, and Markus Krötzsch. 2014. 'Wikidata: A Free Collaborative Knowledgebase'. *Communications of the ACM* 57 (10): 78–85. <https://doi.org/10/gftnsk>.
- W3C Incubator Group. 2011. 'Library Linked Data Incubator Group Final Report'. <http://www.w3.org/2005/Incubator/lld/XGR-ll-d-20111025/>.

“Discoverability” in the IIF digital ecosystem

Paola Manoni^(a)

a) Biblioteca Apostolica Vaticana, <https://orcid.org/0000-0001-7802-2718>

Contact: Paola Manoni, manoni@vatlib.it

Received: 25 August 2021; **Accepted:** 21 September 2021; **First Published:** 15 January 2022

ABSTRACT

The IIF APIs have been used since 2012 by a community of research, national and state libraries, museums, companies and image repositories committed to providing access to image resources. The IIF technical groups have developed compelling tools for the display of more than a billion IIF-compatible images.

We can figure out that with hundreds of institutions participating worldwide, the possibilities, for instance, for IIF-based scholarship are growing so one question could be about the discovery of those images relevant to one's research interests in order to discover them for their consultation or, even more, for their reuse.

While IIF specifications discussion has focused on the machine-to-machine mechanisms of making IIF resources harvestable, we have yet to implement an end-to-end solution that demonstrates how discovery might be accomplished at scale and across a range of differing standards for metadata arising from libraries, archives, and museums.

KEYWORDS

IIF; LAM; Discovery; Digital ecosystem.

1. Discoverability

The International Image Interoperability Framework¹ as an interoperability protocol for image resources held in libraries, archives, museums, has produced over a billion IIIF-compliant images. This paper will focus on how this vast production is actually changing not only the use of digital objects online in the context of tools at the convenience of digital humanities, for instance with the well known abilities of IIIF viewers, such as Mirador², but also the concept of discoverability of the knowledge objects now available via IIIF.

Discoverability is the quality of being able to be discovered or found and in relation to online content, it is the quality of being easy to find via a search engine, within an application, or on a website.

If we focus on the discoverability in the IIIF context we can refer to two main aspects:

- Which are the requirements that make a web platform a discoverable digital library service in the light of IIIF;
- How is it possible to discover IIIF-compliant content through current web platforms.

A first look about the context of the two issues, is concerning the non-trivial meaning of LAM data in the universe of a single domain, for an evaluation of their impact on the discoverability of IIIF objects. We thus consider the abstraction of LAM data produced within the digital ecosystem starting from the traditional statements related to the classes of:

- *Structured data* – In the LAM domain they include bibliographies, catalogs, indexing and abstracting databases, authority files. Structured data is generally stored in databases where all key / value pairs have clear identifiers and relationships and follow an explicit data model
- *Semi-structured data* - they are the unstructured sections within metadata descriptions as well as any unstructured portions of structured datasets.
- *Unstructured data* – they are the typical “everything else” pertaining to documents and other information-bearing objects in all kinds of formats.” (Zeng 2019).

We consider the typical elements of the IIIF Presentation API, keeping in mind that IIIF Presentation API provides:

- A model for describing digital representations of objects: just the metadata chosen in a completely arbitrary way in order to offer a remote viewing experience.
- A format for software - viewing tools, annotation clients, web sites - to consume and render the objects and any other associated content in the form of annotations.

This does not mean that descriptive metadata has no place in a digital object provided by the Presentation API. In fact, it is important that the object is linked to its description and to all the information relating to it. The presentation API provides this human readable information, so that viewers can interpret the important contextual information to end-users.

¹ Cfr. *International Image Interoperability Protocol* < <https://iiif.io/>>. Accessed April 15, 2021.

² Cfr. <https://projectmirador.org/>. Accessed April 15, 2021. Mirador is a fully IIIF-compatible tool capable of interpreting IIIF APIs. Mirador is an open source image, Javascript and HTML5 viewer that delivers high resolution images in a workspace that enables image annotation and comparison of images from repositories dispersed around the world, starting from compatibility with Image API that specifies a web service returning an image in response to a standard HTTP or HTTPS request.

The information pertaining to Presentation API is the IIIF manifest of the digital object represented as a “thing” and enriched with the complex knowledge data related to it.

A so-called IIIF manifest contains a descriptive section of the digital object but the specifications do not define any rules relating to metadata. In other words, we can say that IIIF requirements are completely agnostic as to which descriptive metadata to apply as well as to which image formats. The galaxy of data pertaining to the “thing” represented in the manifest is completely scalable and referable to the different meanings of LAM data (structured, semi-structured, and unstructured) we mentioned.

Moreover, as for the Presentation API, the meaning of any accompanying descriptive metadata for display in a viewer is not taken into any account. The purpose of this API is the representation of the content of the work – for example the pages of the book, the painting – or the link where users can get information about the meaning of the content of the work.

The objective of the IIIF Presentation API is to provide the information necessary to allow a rich, online viewing environment for primarily image-based objects to be presented to a human user, likely in conjunction with the IIIF Image API. In other words, the IIIF Presentation API gives us a specification for “presenting” a digital object and the data describing it in order to view, annotate it, or compare it with other objects. A IIIF client can also display any accompanying metadata included as pairs of labels and values within the manifest. But it needs no definition or scheme for what that metadata means. It is *outside of the scope* of the Presentation API (Crane 2017).

The user can view important semantic metadata, but the scope of the Presentation API is just to leverage that text. In the Presentation API, the semantic meaning is *elsewhere* because it is not belonging to its the specifications. An API client should simply render them.

In a nutshell, a manifest is what a IIIF viewer loads to display the object. A manifest could be used to represent the object within a web service as well as it could be used to add annotations to the represented object or even to be aggregated within a new manifest thus realizing its reuse.

The structure of the manifest also includes the concepts of sequence, which is of fundamental importance for aggregated resources (e.g. books, manuscripts and archive materials composed of page, leaf, folio or sheet) and canvas.

Each view of the object, for example each page is represented by a canvas. A Manifest contains one or more **Sequences of Canvases**. But a canvas is not the same as an image. “The canvas is an abstraction, a virtual container for content” (Crane 2017).

A Canvas is the digital surrogate for a physical page which should be rendered to the user. Each Canvas has a rectangular aspect ratio, and is positioned such that the top left hand corner of the Canvas corresponds to the top left hand corner of a rectangular bounding box around the page, and similarly for the bottom right hand corners. The identifier for the Canvas is not an identifier for the physical page, it identifies the digital representation of it.³

The canvas is a kind of conceptual extra layer in which an object is included.

The canvas keeps the content separate from the conceptual model of the page of the book, paint-

³ Cfr. *Shared Canvas Data Model* <<http://iiif-io.us-east-1.elasticbeanstalk.com/model/shared-canvas/>>. Accessed April 15, 2021.

ing or archival unit. The content, we are referring to, could be blocks of text, videos, links to other resources, and it is exactly mapped on the canvas. By including a canvas in a manifest, you provide a space on which users and scholars can annotate the content.

All association of content with a canvas is done by **annotation**. The IIIF Presentation API is built on the W3C Web Annotation Data Model⁴.

Annotations associate content resources with Canvases. The same mechanism is used for the visible and/or audible resources as is used for transcriptions, commentary, tags and other content. This provides a single, unified method for aligning information, and provides a standards-based framework for distinguishing parts of resources and parts of Canvases.⁵

The canvas establishes a stage in which the simplest case – one image per canvas – is straightforward, but more complex cases, more complex and interesting associations of content, can be managed.

The latest specification of the IIIF, still in beta version, is the IIIF Content State API⁶ which demonstrates another purpose for a representation by sharing a content to be represented on a canvas.

In its scope there are two examples:

- A user follows a link from a search result, which opens a IIIF viewer. The viewer focuses on the relevant part of the object, such as a particular line of text that contains the searched-for term.
- A user opens several IIIF Manifests to compare paintings, then wishes to share this set of views with a colleague.

These are examples of sharing a resource, or better, a *particular view* of a resource. Other examples include bookmarks, citations, playlists and deep linking into digital objects.

The objective of the IIIF Content State API is to provide a standardized format for sharing of a particular view of one or more IIIF Presentation API resources, such as a Collection, a Manifest, or a particular part of a Manifest.

Content State API is **how we can point at things in IIIF** and this demonstrates how the concept of digital resource and its reuse expand to include new ways of knowing the resources and new ways of citing them. In fact, it basically means dereferencing URIs of annotations whose motivation such as *content state* will be included in a manifest.

Content State is a way for humans to share bookmarks, and it's also a way for search results to point at the exact part of a digital object that they match (Crane 2021).

We can argue at this regard that the semantic enrichment process pertaining to the IIIF's vision of LAM objects and data reflects the broader general transformation from document-centric to entity-centric knowledge modeling due to the many relations for each canvas.

⁴ Cfr. *Web Annotation Data Model*. Accessed April 15, 2021. <https://www.w3.org/TR/annotation-model/>. The Model does not prescribe a transport protocol for creating, managing and retrieving annotations. Instead, it describes a resource oriented structure and serialization of that structure that could be carried over many different protocols.

⁵ Cfr. *IIIF Presentation API 3.0* Accessed April 15, 2021. <https://iiif.io/api/presentation/3.0/>.

⁶ Cfr. *IIIF Content State API 0.3* Accessed April 15, 2021. <https://iiif.io/api/content-state/0.3/>.

Let us now go back to consider the discoverability of the IIIF in the light of Presentation API and the first question

- Which are the requirements that make a web platform a discoverable digital library service in the light of the IIIF;

We may focus on this by considering the use case of the Vatican Library as an example.

2. The use case of the Vatican Library

DVL (the DigiVatLib, <<https://digi.vatlib.it>>) is a digital library service. It provides free access to the Vatican Library's digitized collections: manuscripts, incunabula, archival materials and inventories as well as graphic materials, coins and medals, printed materials. It is fully based on the International Image Interoperability Framework technology, making digital materials easily accessible and usable.

- The viewer is able to zoom, browse and 'turn pages' of JPEG2000 images as well as allow scholars to compare digital objects from different IIIF repositories of other digital libraries.
- Descriptions and bibliographic references from the online catalogues are indexed and linked to digital materials.
- Each object is equipped with URIs for the discovery of IIIF manifests.
- The guided navigation ('faceted search') leverages metadata elements for narrowing or refining queries.

The Library has promoted a *new* perspective to the study of manuscripts by means of web communication and IIIF.

To meet this challenge the Library has implemented a project to enrich the digital delivery of these materials by annotating some exemplary manuscripts with scholarly analysis.

The use case of annotations in IIIF was a three-year Mellon-funded project, held between 2016 and 2019, in conjunction with Stanford University Libraries, which produced over 26,000 annotations for a selection of manuscripts chosen in the context of thematic pathways. In this platform (available at: <<https://spotlight.vatlib.it>>) the content of all the annotations is indexed along with the metadata, thus constituting a semantically enriched system that allows scholars to query an integrated search of all the available contents of a resource.

The project aimed to demonstrate, among the advantages of the IIIF for manuscripts, how the annotation level is a fundamental innovation for the study of contents: transcriptions, comments, comparative analysis of texts and images.

Thanks to the funds received, the Library has implemented a workflow using Mirador with scholarly analysis in order to tell scholarly narratives.

The Vatican Library has intended to engage the visitors to its website on the possibilities for using annotated manuscripts in IIIF, according to specific themes, by providing tools for discovering and comparing digital materials.

The deep analysis of contents of manuscripts entails the understanding of the "pre-print" world in which the manuscript is born. This implies a knowledge pertaining to the history of the man-

uscript, its origin, provenance as well as other circumstances of the production of a manuscript; identifications of dates, scribes, artists; discussions about the intellectual content and descriptive discussion on paleographic matters.

In its essential lines, a thematic pathway is composed by three different kinds of information:

- A general description (introduction, historical information, etc) of the chosen theme, it represents the “Story”;
- Descriptive and structural metadata and a curatorial narratives for each manuscript;
- Annotations, comments, in-depth analysis about detailed parts of a manuscript (e.g. texts, comments, illuminations, etc.) and transcriptions of units of information.

The four thematic pathways

1. The first one is about *Courses in Paleography (Greek and Latin, from antiquity to the Renaissance)*

The rich collection of manuscripts preserved in the Library makes it possible to follow the evolution of the Greek and Latin scripts all the way from antiquity to the Renaissance.

The availability of on-line images of manuscripts, together with the possibilities offered by the IIF APIs, allows a complete transformation of teaching practice in this field.

For each of the sections (Greek and Latin) of this thematic path, a set of complete digitized manuscripts, chosen to illustrate the phases in the development of the script from the fourth to the sixteenth century, is provided. From each manuscript, chosen pages with a paleographical and codicological description and a diplomatic transcription is also made available.

2. The second one is about *The evolution and transmission of texts of specific works: Latin Classics*

The Vatican Library owns one of the most important collections of manuscripts with texts by Classical Latin authors, many of them richly illustrated.

The aim of this pathways is to describe 81 manuscripts directly from the original codices: metadata and annotations pertaining to the study of texts and illuminations have been provided. The work throws light not only on the illustrations of the texts but especially on the relationship between text, illuminations, comment and the gloss.

The importance of this project lies in the remarkable variety of typologies of the Classical world.

3. The third one is about *Vatican Palimpsests: Digital Recovery of Erased Identities*

The Vatican Library has identified more than 380 manuscripts in its own collections, which include palimpsests, erased and then recycled parchment folios. This pathway intends to present this rich and scarcely explored material to the public by making an in-depth archaeological research on the palimpsests of twenty-four select manuscripts and recover their lost identities with the help of IIF technology.

Making accessible hardly legible images to the public is a challenging task because the

actual method of publication has been designed to typical objects. By the pathway, digital reconstruction makes four palimpsests accessible both by their upper and lower scripts, a condition which the actual conservation of these manuscripts and the normal method of publication do not allow.

Erased texts are often very old and significant witnesses of a lost past but they are difficult to access for the naked eye. They need an expert interpreter and highly special photographic and post-processing technologies and especially the flexibility of presentation offered by IIIF APIs which can turn erased texts more accessible online than in their physical existence.

4. The last one is about The humanist prince's library: Federico da Montefeltro and his manuscripts

The library of Federico da Montefeltro, Duke of Urbino (since 1474), is known as a typical humanist collection.

The collection was outstanding not only for its substance (the amount of volumes as well as the quality, in relation with other libraries of that age), but for the value of each manuscript partly acquired from antique market, many commissioned by Federico and realized by refined copyists and greater artists of that time. The manuscripts were produced in two main locations: Florence and Urbino.

In the first years, Federico preferred to buy or order manuscripts in Florence (both in writing and in illumination), later he preferred Ferrara or Padoan artists and scribes active in Urbino.

This pathway points out the characteristics of the two schools, very different in style, and the most important artists (half of the chosen manuscripts is representative of the Florentine school while the other half of the Ferrara and Padoan schools).

3. IIIF Discovery for Humans Community Group

IIIF enables the creation of rich digital collections that bring together content distributed among cultural heritage institutions. With image viewers, one is able to analyze works held in physically different locations side-by-side or overlaid within a web browser. However, in order to take advantage of the research tools afforded by IIIF, a user must be able to find IIIF resources.

Interoperable objects are of no use if one cannot find them, particularly if relevant objects reside in servers in many different institutions. Discovery in this case means human searching, browsing and finding of IIIF resources across institutions. To be successful, IIIF discovery must be user-focused and meet defined users' concrete needs.

To meet these needs, the IIIF Discovery for Humans Community Group was recently organized. This group aims to go beyond specification work to promote implementations that enroll experts in research, content, user experience, metadata, and various technologies. In order to advance discovery in the LAM space, this group will foster user-focused approaches enabling the targeted discovery, spanning institutional and domain silos, of IIIF resources.

These aims are different from and complementary to the approach of the IIIF Discovery Tech-

nical Specification Group, which is chiefly concerned about providing the technical means for locating and finding updates about IIIF resources, as a prerequisite for harvesting and indexing metadata for searching within and across these institutional collections.

If we focus again on the two questions arisen in this paper about:

- Which are the requirements that make a web platform a discoverable digital library service in the light of the IIIF;
- How is it possible to discover IIIF-compliant content through current web platforms.

We may say that both are of fundamental interest to this group and they are closely related to the purposes of the initiatives conducted by the Group, aimed at:

- Gather problem statements and use cases to understand needs for user-focused discovery of IIIF resources
- Develop specifications for metadata attributes and crosswalks to enable discovery of LAM IIIF content across institutions and domains
- Create and maintain a list of metadata profiles in use by IIIF-supporting institutions to promote consistency in semantic description and its consumption
- Frame small-scale experiments that work towards live discovery implementations
- Provide a venue for demonstrations of applicable discovery applications and technologies
- Maintain a registry of existing discovery efforts
- Build on and amplify the ongoing work of the Discovery Technical Specifications Group and the IIIF Technical Community
- Communicate and disseminate the work of the group to the larger IIIF community, as well as allied professional communities.

One of the recent activities of this group was to **collect a list of discovery features** in order to:

- Provide examples of features as implemented
- Extract a comprehensive list of discovery features
- Develop a feature typology
- Identify IIIF-specific discovery affordances
- Build a feature checklist for self-evaluation
- Inform development of other discovery platforms
- Showcase exemplary feature implementations

First of all this task has provided a collated list of discovery features “in the wild,” to better understand the current landscape of IIIF resource discovery. It was useful as a basis for compiling a wide-ranging list of possible discovery features. This was further condensed and organized to derive broader categories for these features, and to identify which features were specifically tied to IIIF rather than associated discovery more generally. From this analysis we were able to get a sense of which features are broadly implemented and which are rarer. Based on this work we developed a feature checklist that may be used to identify which core discovery features are and are not available on a given site.

Metric for discovery features was the first important milestone planning the group’s commitments as an important first step to face the second question: How is it possible to discover IIIF-compliant content through current web platforms, the thread underlying this paper.

References

- Crane, Tom. 2017. *An Introduction to IIIF*. Digirati. Accessed April, 15 2021. <http://resources.digirati.com/iiif/an-introduction-to-iiif/>.
- Crane, Tom. 2021. *What is IIIF Content State?* Accessed April 15 2021. <https://tom-crane.medium.com/what-is-iiif-content-state-dd15a543939f>.
- Manoni, Paola. 2020. "L'adozione del IIIF nell'ecosistema digitale della Biblioteca Apostolica Vaticana." *DigItalia* 2: 96-105.
- Manoni, Paola - Ponzi, Eva. 2020. "Thematic Pathways on the Web: IIIF Annotations of Manuscripts from the Vatican Collections: il "Progetto Mellon" della Biblioteca Vaticana". *Rivista di Storia della Miniatura* 24: 211-216.
- Salarelli, Alberto. 2017. "International Image Interoperability Framework (IIIF): a panoramic view". *JLIS* 8, no. 1. Accessed April,15 2021. doi: 10.4403/jlis.it-12090.
- Zeng, Marcia Lei. 2019. "Semantic enrichment for enhancing LAM data and supporting digital humanities: Review article" *El profesional de la información* 28, no 1.

Bibliographic Control of Research Datasets: reflections from the EUI Library*

Thomas Bourke^(a)

a) European University Institute

Contact: Thomas Bourke, thomas.bourke@eui.eu

Received: 7 April 2021; **Accepted:** 3 June 2021; **First Published:** 15 January 2022

ABSTRACT

The exponential growth in the generation and use of research data has important consequences for scientific culture and library mandates. This paper explores how the bibliographic control function in one academic library has been expanded to embrace research data in the social sciences and humanities. Library bibliographic control (BC) of research datasets has emerged at the same time as library research data management (RDM). These two functions are driven by digital change; the rise of the open science and open data movements; library management of institutional repositories; and the increasing recognition that data sharing serves the advancement of science, the economy and society. Both the research data management function and the bibliographic control function can be enhanced by librarians' awareness of scholarly projects throughout the research data lifecycle (input, elaboration and output) – and not only when research datasets are submitted for deposit. These library roles require knowledge of data sources and provenance; research project context; database copyright; data protection; data documentation and the FAIR Guiding Principles, to make data findable, accessible, interoperable and reusable. This case study suggests that by creating synergies between the research data management function (during research projects) and the formal bibliographic control function (at the end of research projects) – librarians can make an enhanced contribution to good scientific practice and responsible research.

KEYWORDS

Research data | Datasets | Research data management | Bibliographic control.

* With special thanks, for comments and contributions, to Carlotta Alpigiano (EUI Acquisitions and Library Budget, Co-ordinator), Tommaso Giordano (Former Director, EUI Library), Simone Sacchi (EUI Open Science Librarian), Monica Steletti (EUI Special Collections Librarian), Lotta Svantesson (EUI Repository Manager) and Pep Torn (EUI Library Director). Thomas Bourke is EUI Library information specialist for economics.

© 2022, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.



1. Introduction

Research libraries have a long history of collecting, managing and providing access to data resources – in particular, statistical data series in support of the social sciences. While there are important differences between disciplines and sub-disciplines, most research libraries had a limited role in the management of their institutions' research data *outputs* until the 21st Century. Today, the collation, bibliographic control, preservation and dissemination of research data are important library functions, due to increasing awareness of datasets as 'first-class' outputs of research.

This case study treats the management and bibliographic control of research dataset outputs in the social sciences and humanities at the Library of the European University Institute (EUI).

While research data management (RDM) is primarily carried out by scholars during research projects, librarians have steadily increased their collaboration and training to fill the skills and capabilities' gap in this area. RDM is undertaken throughout the research data lifecycle, embracing the control of data inputs, the elaboration of data, the protection of data and the creation of research data outputs and documentation. The main reasons for librarians' involvement in research data management include: the exponential growth in the availability and production of digital data; the rise of the open science and open data movements; the establishment of research repositories – frequently managed by libraries; and the increasing recognition that the sharing of research data serves the advancement of science. All of the above are in contexts which require the transfer of knowledge and expertise of library staff – through consultation with, and training of, researchers. While librarians' research data management takes place *during* research projects; bibliographic control normally takes place *after* (or towards the end of) research projects.

Both research data management (Section 3 below) and the bibliographic control of datasets (Section 4 below) require librarians to strengthen their liaison with researchers in order to enhance familiarity with research project design, data generation and use, and research data outputs. Research data management also requires librarians to support the generation of data management plans (DMPs) which are increasingly required by science funders.¹ Support for data management planning raises librarians' awareness of the nature and scope of research projects before final data outputs are presented for deposit. While campus libraries have important roles regarding research data management and bibliographic control, it is acknowledged that – in some institutions – there are lead roles for data centres, ICT services and/or research administration offices.

2. Data, digital change and scientific culture

The generation, collection and use of vast quantities of data – and the retro-digitisation of non-digital collections and content – places research libraries at the vanguard of recent transformation.² In

¹ The European context is described by Filip Kruse and Jesper Boserup Thestrup (Kruse and Thestrup 2018).

² The evolution of library research data roles is analysed by Robin Rice and John Southall (Rice and Southall 2016); by Lynda Kellam and Kristi Thompson, et al. (Kellam and Thompson 2016); and by Rossana Morriello (Morriello 2020).

addition to the volume of data, the tools for the elaboration of data have become more sophisticated. In the social sciences and the humanities, these developments have had an impact on scientific culture – facilitating more empirical and applied research; experimental research; evidence-based policy research, and data-driven methodologies.

The definition of ‘data’ varies across academic disciplines and sub-disciplines and the scope of the term itself has been debated for several decades.³ Data types in the social sciences and humanities include: numerical data, minable text, survey data, experimental data, interview transcripts, archival material, field notes, images, and audio and video recordings. The long history of library expertise in the management of multi-media collections constitutes a solid basis for library curation of research data outputs.⁴

Although definitions vary by discipline, it is useful for librarians to distinguish between collected or acquired ‘databases’ (eg. databases of monographic, journal or statistical content) and individual ‘datasets’ (eg. data outputs from research projects hosted at their institutions). This paper does not treat the traditional library database acquisition and management function (which includes the acquisition, classification, cataloguing and access control of subscription resources; eg. financial market data).⁵ The data treated in this case study are *outputs* generated by university scholars, which are managed, bibliographically controlled, curated, classified and repositied by library staff for the purpose of preservation and – where possible – sharing with other researchers.

The open data movement – an extension of the open access movement – refers to a growing trend whereby government agencies, international organisations and researchers share data outputs, documentation, codebooks and software via the internet. Here it is necessary to distinguish between ‘public data’ and ‘research data.’ Most governments and international organisations provide some level of access to ‘open public data.’ In the research community ‘open research data’ refers to outputs from scholarly research projects which are openly available, usually via institutional repositories.

3. Research data management: library roles

The impact of technological change on scientific culture has necessitated the expansion of library data support roles. The traditional function of acquiring access to subscription databases has been joined by two newer library roles: (i) library support for data-intensive research and research data management *during the research data lifecycle* and (ii) the bibliographic control, collation, reposit and preservation of research data outputs *at the end of the research project*. Research data management during research projects is carried out by both scholars and librarians, complementing their

³ For a theoretical treatment, see the entry “Data” in the Encyclopaedia of Knowledge Organization: <https://www.isko.org/cyclo/data> (International Society for Knowledge Organization, n.d.).

⁴ See Joudrey 2015, chap. 5; and Pradhan 2018.

⁵ At the EUI, these resources are presented in the Library Data Portal: <https://www.eui.eu/Research/Library/Research-Guides/Economics/Statistics/DataPortal> and classified at: <https://www.eui.eu/Documents/Research/Library/Research-Guides/Economics/Statistics/MacroMicroLocations.xls> (Accessed 7 April 2021).

respective expertise.⁶ Bibliographic control, reposit and preservation of research data outputs are carried out by librarians. This is especially true when it comes to datasets in disciplines where the culture of managing (and sharing) research data is not yet fully developed, or where established subject-oriented data repositories (e.g. GenBank, HEPData) do not exist.

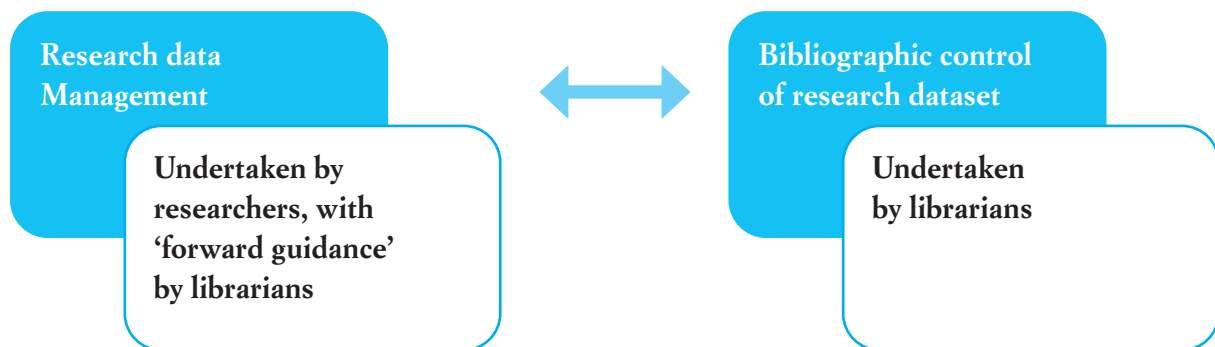


Fig. 1. RDM and BC roles: scholars and librarians

An important component of research data management is the generation of data management plans (DMPs).⁷ A request from a principal investigator for DMP support is often the first point of contact between a librarian and a new research project. Data management plans provide information on how data is generated and/or sourced; how data is organised, used and elaborated; how data – and data subjects – are protected; how data and tools are described and documented; how data is stored and secured during the research project; how data authorship and credit are assigned; how data will be preserved and whether research data outputs can be shared.⁸ The involvement of librarians in data management planning constitutes a solid foundation for the eventual bibliographic control of dataset outputs.

Contemporary research data management is underpinned by the FAIR Guiding Principles, to make data *findable*, *accessible*, *interoperable* and *reusable* – frequently used by librarians to promote awareness of good research data management practices.⁹ Both the research data management function and the bibliographic control function help advance the FAIR Guiding Principles. Library research data management 'forward guidance' is provided via individual user support, library-web documentation and group training. Librarians provide advice that data outputs must be carefully structured, because the 'objects' (outputs) for reposit will be datasets (not unstructured data observations) and that researchers should carefully consider the design of datasets early in their research projects. Dataset structure varies by discipline and sub-discipline, medium, types

⁶ The Consortium of European Social Science Data Archives (CESSDA) maintains an annually-updated Data Management Expert Guide. <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide> (Consortium of European Social Science Data Archives, n.d.).

⁷ Online template tools such as DMPonline <https://dmponline.dcc.ac.uk/>, maintained by the UK Digital Curation Centre (Digital Curation Centre, n.d.), and Argos <https://argos.openaire.eu/splash/>, maintained by OpenAIRE (OpenAIRE, n.d.), can be used to generate structured data management plans.

⁸ <https://www.eui.eu/Research/Library/ResearchDataServices/Guide> (European University Institute Library 2021).

⁹ See Wilkinson, Dumontier, Aalbersberg et al. 2016. Barend Mons provides a practical overview (Mons 2018).

of variables, units of analysis, relationships between data elements, and whether or not the dataset is part of a series. Librarians also explain the importance of clear and consistent naming of folders, files, variables, versioning and documentation, and how good practice helps facilitate findability, accessibility, interoperability and reusability.

Supporting documentation should be updated throughout research projects because, when datasets are presented for deposit, documentation – such as codebooks and questionnaires – must also be submitted. Comprehensive documentation – describing dataset structure, folders, files, variables, versioning and (where applicable) information about problematic values, missing observations and weightings – makes research data findable, accessible, interoperable and re-usable (FAIR). Librarians – who are familiar with a wide variety of data documentation across disciplines and sub-disciplines – can offer feedback on data documentation and help edit dataset abstracts at the time of deposit.

Although all research institutions have data protection officers (DPOs), the library is frequently the first point of contact for scholars who have questions about database copyright and data protection. Librarians' long-standing experience with copyright and terms and conditions of access and use, has been extended to database copyright – which is important when library-licensed databases are used by researchers to generate new research data outputs. In many social science research projects, data outputs are the product of 'mixing' pre-existing data resources (frequently acquired and made available by the campus library) with new project-generated data (eg. surveys and experiments). For librarians, who are 'custodians' of subscription databases, it is important to inform database users of terms and conditions of access and use before research datasets based on licensed resources are openly shared.

During the research data lifecycle – and in collaboration with the DPO – librarians also inform scholars of their data protection obligations regarding the collection, use and security of data observations relating to persons, families and households. Librarians can advise on anonymisation and pseudonymisation techniques – which are particularly relevant for micro-level socio-economic data.

At the library of the European University Institute it is observed that many of the features of research data management (RDM) during the research project overlap with – and help prepare for – formal bibliographic control (BC) at the end – or near the end – of research projects.

4. Bibliographic control, infrastructure and workflow

Due to technological change, the exponential increase in digital content, and the momentum of the open access movement – universities began to establish institutional research repositories at the turn of the 21st Century.¹⁰ Initially these infrastructures only indexed full-text documents and bibliographic records of publications. Gradually research dataset outputs and multi-media have been added, due to an ongoing research culture change towards open science and the increasing requirements of funding agencies. Research scholars also have the option to deposit

¹⁰ Data on the growth of repositories (2000-2020) is available from the Registry of Open Access Repositories: https://en.wikipedia.org/wiki/Registry_of_Open_Access_Repositories. Accessed 7 April 2021.

their data outputs in subject/domain repositories and ‘catch-all’ multi-disciplinary repositories, such as Zenodo.¹¹

The EUI Library launched the Cadmus institutional repository, based on the DSpace infrastructure, in 2003.¹² The beta version of the EUI ResData repository was launched in 2016 and was merged with Cadmus in 2019. University librarians are increasingly aware of data-driven research projects because the campus library is the primary source for subscription databases; researchers usually require access to data software manuals provided by the library; data management plans are supported and reviewed by librarians and researchers frequently approach the library for advice on database copyright and data protection during the research data lifecycle.

However, it is not possible for librarians to be aware of every data-intensive project on campus, as there is no mandate for such information to be shared. Sometimes, librarians will only become aware of research data outputs when a principal investigator, or research team, approaches the library for advice on the preservation, reposit and open sharing of research dataset outputs – often due to funding agency requirements, such as the Horizon 2020 Framework Programme Open Research Data Pilot. Figure 2 provides an overview of the roles of researchers, librarians and ICT staff at the EUI during the research data lifecycle.

	ACTIVITY	RESEARCHERS	LIBRARY	ICT Service
Data input	Data discovery	Researchers discover data via library collections; the internet; and non-digital resources	Maintenance of data portal, indices and OPAC records	/
	Data generation	Researchers generate data (eg. surveys, experiments)	/	/
	Terms of access and use; database copyright and data protection	User compliance	Library promotes awareness of terms and conditions of access and use	ICT service and library provide access protocols
	Data management plans (DMPs)	Researchers write data management plans	Library provides training on DMP template tools and helps edit DMPs	ICT service provides standard description of infrastructure and security



¹¹ <https://zenodo.org/>. Accessed 7 April 2021.

¹² <https://cadmus.eui.eu/>. Accessed 7 April 2021.

	ACTIVITY	RESEARCHERS	LIBRARY	ICT Service
Data elaboration / In-project data management	Dataset structure: folders, files, variables, observations	Researcher activity	Library advisory role	/
	Data anonymisation	Researcher activity	Library advisory role	
	Standardisation of file names, versioning, in-project metadata	Researcher activity	Library advisory role	/
	Documentation, codebooks and associated software/ routines	Researcher activity	Library advisory role	ICT advisory role
	In-project security and backup	Researcher activity	Library advisory role	ICT infrastructure and encryption software
Data output	Submitting research datasets	Researchers submit details of data outputs via online form	Library reviews submission	/
	Bibliographic control and metadata	/	Library checks structure and sources of dataset; converts submission information into metadata	/
	Repositing and infrastructure	/	Library reposit datasets in the institutional repository	Support for institutional repository infrastructure

Fig. 2. Research data lifecycle roles of researchers, librarians and ICT staff

4.1 Data submission to the institutional research repository

Over the past two decades there has been a growing awareness of the scientific value of making research data more openly available. While many academic disciplines have a long history of sharing underlying data within epistemic communities – the open science movement advocates wider access to research data as a public good, of benefit to scientific endeavour. Researchers in the social sciences and humanities are increasingly aware that the academic community is awarding recognition to research datasets as outputs in their own right. Reposited datasets can promote awareness of related publications, or in themselves become part of promotion and tenure procedures. In some cases, researchers become aware of these issues (open data; open science) late in the research project – for example, when an academic colleague or a funding agency requests information about underlying data. Researchers at the European University Institute who submit datasets for reposit are required to complete the library’s online data submission form.¹³ It is important to distinguish between three

¹³ <https://www.eui.eu/Research/Library/ResearchDataServices/EUIResDataWorkflow>. Accessed 7 April 2021.

types of data description activities. Firstly, researchers generate essential descriptors for their data (names of folders, files, tabs, variables &c.) during the research project. These descriptors do not always constitute formal ‘metadata’ in the sense of bibliographic control. Secondly, the observations entered by researchers in the EUI library’s online data submission form constitute ‘raw’ information about a dataset, and are never ingested directly into the repository without review. Thirdly, librarians generate bibliographic-standard metadata for the research repository; to make research datasets findable, accessible, interoperable and reusable.

This case study suggests that the in-project research data management (RDM) function complements the end-of-project bibliographic control (BC) function. The creation of synergies between the two library functions helps contribute to overall scientific quality control.

4.2 Initial review

The EUI library’s online data submission form captures information which is used for verification, provenance and bibliographic control. At the EUI, the name and institutional email address of the principal investigator (or delegated submitter) is required for verification that the submitter is a member of the institution. Only works generated by EUI members – or research teams with at least one EUI member – can be included in the institutional repository. Alumni and former professors can submit datasets if the substantive part of the research was conducted while a full member of the university.

The EUI library became a member of ORCID, the Open Researcher and Contributor ID service, in November 2017. ORCID is a solution for authority control of authors’ name variations across the EUI’s Central Person Registry (CPR); the research repository Cadmus, and the ORCID registry. Both publications and datasets are associated with authors’ ORCID IDs, providing increased visibility for researchers and the institution in the digital environment. EUI authors’ names in the Cadmus repository are linked to the ORCID record – pushing publication and dataset metadata to their ORCID profiles.¹⁴

- When completing the dataset submission form, the names of all creators of the dataset must be listed – including technical collaborators if they have significantly contributed to the creation of the dataset.
- The title of the dataset submitted should not be identical to the title of a project or a publication. Librarians frequently offer suggestions regarding title clarity and, in many cases, titles are modified.
- The online submission form captures both the year of completion of the dataset (which may, or may not, be the current year); the date-range of data coverage – which is of great importance for any time-series data; and (where applicable) the geographical coverage of the dataset.
- Submitters are required to provide a description of the dataset – which is a first draft of the abstract displayed prominently in the repository entry.
- One of the most important submission form fields is the ‘Source(s) of data’. If the dataset is the output of original data collection and elaboration, details should be provided. If the

¹⁴ The complete workflow is explained by Lotta Svantesson and Monica Steletti, in their presentation at the Open Repositories conference (Svantesson and Steletti 2019). The EUI ORCID connect page is at: <https://cadmus.eui.eu/ORCID/>. Accessed 7 April 2021.

dataset is derived from pre-existing sources, those sources should be clearly indicated (data creator, institutional source, publisher).

- The online submission form requires a preliminary statement of whether the data can be made available for open sharing immediately, or is to be reposited under embargo. Librarians help determine this status in consultation with researchers.
- Submitters are required to provide the file format of data files. If the data is in a proprietary format, librarians can recommend (where possible) options for open format versions. This information is always translated into the related media type.¹⁵
- The number of data files within the dataset is an important field which allows librarians to discuss the relationship between the repository entry and the constituent elements. In some cases, it is necessary to create two entries (works) for a dataset which the submitter might be submitting as a single work. In other cases, multiple data submissions from the same project might be consolidated into one entry with multiple sub-sets.
- The online form also requests information regarding projected future waves of the dataset being submitted. This can require that a data sub-set which is intended to have future iterations, might need a separate entry in the repository.
- The online form also gathers information about supporting documentation, codebooks and (where applicable) software routines to enable the use of the data by others.
- The library advises on the appropriate reuse licence for open research data; eg: Creative Commons Attribution (CC-BY) or Public Domain (CC0).
- Submitters are asked to include references to related publications. This information can also be added when publications become available.

4.3 Provenance

The establishment of research repositories in universities and other institutions requires librarians to have a strong role regarding provenance. While research documents (working papers, theses, articles, chapters, monographs &c.) are normally subject to editorial review either inside the university or by external peer reviewers and publishers – the situation regarding research data outputs is more complex.

Very few universities have formal faculty-level ‘editorial’ review procedures for dataset outputs. The research data lifecycle is predominantly undertaken by researchers – with support from library and ICT professionals. At the end of research projects, the library becomes involved in issues of provenance, originality, data protection and database copyright. Here it can be seen that there is an overlap between the research data management (RDM) function and the bibliographic control (BC) function.

Although there are multiple ways in which librarians can undertake verification and provenance, it is impossible for librarians and information specialists to have detailed knowledge of data in every discipline and sub-discipline. It is also impossible for librarians to guarantee that every element and observation in a dataset is correct.

¹⁵ Formerly known as MIME Type: https://en.wikipedia.org/wiki/Media_type. Accessed 7 April 2021.

At the EUI, librarians build trust with researchers during the data lifecycle as part of the research data management function – informing scholars that;

By submitting this [online submission] form, EUI members acknowledge that the dataset for deposit is the output of original data collection and elaboration; or is the output of significant, value-added, elaboration of pre-existing sources; and conforms with the EUI *Guide to Good Data Protection Practice in Research*.¹⁶

Librarians build trust with researchers through outreach and training; assistance with data management plans; provision of in-project services during the research data lifecycle and advice about database copyright, data protection, research ethics, scholarly reputation and scientific impact. When data is presented to the library for deposit, the ‘Source(s) of data’ field in the online submission form reveals whether the dataset output is partially based on pre-existing, library-licensed resources. At this point, it may be necessary for library staff to liaise with data suppliers to control for potential license issues regarding the open sharing of derivative datasets via the university repository.

4.4 Metadata generation

The generation of metadata about research datasets renders research datasets findable, accessible, interoperable and reusable (FAIR) and helps librarians decide whether research data outputs can be shared as open data. The research data management activities undertaken by librarians during research projects constitute a solid foundation for library bibliographic control and metadata generation. At the EUI, librarians use the raw information from the online dataset submission form to generate repository metadata using:

- The Dublin Core schema
- Library of Congress subject headings
- Dewey Decimal 23 classification
- A modified UN/Eurostat classification originally developed for the paper-format statistics collection¹⁷ and,
- An internal data series identifier.

When setting up the EUI Research Data Collection structure in the Cadmus institutional repository, the EUI’s institutional setup was reflected in the sequential, internal ID (dc.identifier.other) – eg: EUI_ResData_00032_HEC. The numeric value is a running sequence, with alpha-suffixes for:

- Economics: ECO
- History and civilisation: HEC
- Law: LAW
- Social and political sciences: SPS and
- The inter-disciplinary Robert Schuman Centre for Advanced Studies: RSC.

Here follows an example of the metadata record for a dataset deposited in 2020.

¹⁶ <https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf> (European University Institute 2019).

¹⁷ <https://www.eui.eu/Research/Library/ResearchGuides/Economics/StatisticsClassification>. Accessed 7 April 2021.

Informal politics of codecision dataset	
dc.contributor.author	BRESSANELLI, Edoardo
dc.contributor.author	HERITIER, Adrienne
dc.contributor.author	KOOP, Christel
dc.contributor.author	REH, Christine
dc.coverage.spatial	European Union
dc.coverage.temporal	1999-2009
dc.date.accessioned	2020-09-09
dc.date.available	2020-09-09
dc.date.created	2014
dc.date.issued	2020
dc.identifier.other	EUI_ResData_00028_RSC
dc.identifier.uri	https://hdl.handle.net/1814/68095
dc.description	1 data file; 1 documentation file
dc.description.abstract	This dataset, created as part of the research project on ‘The Informal Politics of Codecision’ - funded by the Research Council of the European University Institute (EUI) and the Economic and Social Research Council (ESRC; Grant RES-000-22-3661) - is constituted by all 797 legislative files concluded under codecision between 1999 and 2009. It presents a new variable, ‘early agreement’, indicating whether legislation has been agreed informally, in trilogues, by the Council of Ministers and the European Parliament. It also includes variables with characteristics of the legislative file (legal nature, policy area, complexity, media salience, policy type, duration) and of the legislative negotiators (priorities of the Council Presidency, ideological distance between the Parliament’s rapporteur and the national minister, the Presidency’s workload).
dc.format	Excel file
dc.format.mimetype	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
dc.language.iso	en
dc.publisher	European University Institute, RSCAS
dc.relation.ispartofseries	EUI Research Data
dc.relation.ispartofseries	2020
dc.relation.ispartofseries	Robert Schuman Centre for Advanced Studies
dc.rights	info:eu-repo/semantics/openAccess
dc.rights.uri	http://creativecommons.org/licenses/by/4.0/
dc.subject	Legislative bodies
dc.subject.classification	FS-CA
dc.subject.ddc	328.4077
dc.subject.lcsh	Legislative bodies - European Union countries



dc.title	Informal politics of codecision dataset
dc.type	Dataset
eui.subscribe.skip	TRUE
dc.rights.license	Creative Commons Attribution 4.0 International
dc.description.version	The dataset documentation is available in: BRESSANELLI, Edoardo, HERITIER, Adrienne, KOOP, Christel, REH, Christine, The informal politics of codecision : introducing a new data set on early agreements in the European Union, EUI RSCAS, 2014/64, EUDO - European Union Democracy Observatory -- Retrieved from Cadmus, European University Institute Research Repository, at: http://hdl.handle.net/1814/31612

Fig. 3. Example of metadata record, full view, from the EUI Cadmus repository Research Data Collection

This case study suggests that the in-project research data management (RDM) function complements the end-of-project bibliographic control (BC) function. The generation of metadata about research datasets helps to make research datasets findable, accessible, interoperable and reusable (FAIR). For example, the unique and persistent identifier helps researchers to find the dataset; the retrievability of the metadata via the repository protocol helps make the dataset accessible; the Dublin Core schema allows for broad sharing and interoperability, and the license information facilitates reusability.

4.5 Transfer and uploading of datasets

When EUI librarians have prepared the metadata record for the dataset, an appointment is made for the transfer of data, documentation and (where applicable) codebooks. At this stage, there may be further discussions about structure, format, provenance, copyright and data protection. Once the dataset is received and approved, the metadata file, the dataset and the documentation are uploaded in the Research Data Collection of the EUI Cadmus repository. A digital object identifier is generated and a data citation can be exported, eg:

BRESSANELLI, Edoardo, HERITIER, Adrienne, KOOP, Christel, REH, Christine, *Informal politics of codecision dataset*, EUI Research Data, 2020, Robert Schuman Centre for Advanced Studies. Retrieved from Cadmus, European University Institute Research Repository, at: <https://hdl.handle.net/1814/68095>

The repository metadata schema allows further discovery of the resource, for example via library discovery tools and online aggregator services.¹⁸ Accurate bibliographic control will also facilitate forthcoming machine discoverability of datasets and artificial intelligence applications.

¹⁸ The EUI's Cadmus repository is interoperable with, and harvested by, CORE, Google Scholar, OpenAIRE, RePEC and Worldcat.

5. Conclusion

Contemporary research data management is underpinned by the FAIR Guiding Principles, to make data findable, accessible, interoperable and reusable. Both the research data management (RDM) function and the bibliographic control (BC) function can be combined in the service of these principles.

Based on the experience of EUI library staff – this paper suggests that research data management during research projects and bibliographic control at the end of research projects are complementary elements of an emerging ‘continuum’ of library support for modern scientific culture – contributing to overall scientific quality control.

References

- Consortium of European Social Science Data Archives. n.d. "Data Management Expert Guide." Accessed 7 April 2021. <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide>.
- Digital Curation Centre. n.d. "DMPonline data management tool." Accessed 7 April 2021. <https://dmponline.dcc.ac.uk/>.
- European University Institute. 2019. *Guide to Good Data Protection Practice in Research*. Accessed 7 April 2021. <https://www.eui.eu/documents/servicesadmin/deanofstudies/researchethics/guide-data-protection-research.pdf>.
- European University Institute Library. 2021. *Research Data Guide*. Accessed 7 April 2021. <https://www.eui.eu/Research/Library/ResearchDataServices/Guide>.
- International Society for Knowledge Organization. n.d. "Data". In *Encyclopaedia of Knowledge Organization*. Accessed 7 April 2021. <https://www.isko.org/cyclo/data>.
- Joudrey, Daniel N., Arlene G. Taylor, and David P. Miller. 2015. *Introduction to Cataloging and Classification*. 11th ed. Santa Barbara, CA: Libraries Unlimited.
- Kellam, Lynda, and Kristi Thompson, eds. 2016. "Databrarianship: the Academic Data Librarian" In *Theory and Practice*. Chicago, IL: Association of College and Research Libraries.
- Kruse, Filip, and Jesper Boserup Thestrup, eds. 2018. *Research Data Management: a European Perspective*. Berlin: De Gruyter Saur.
- Mons, Barend. 2018. *Data Stewardship for Open Science: implementing FAIR Principles*. Boca Raton, FL: CRC Press.
- Morriello, Rossana. 2020. "Birth and Development of Data Librarianship." *JLIS.it* 11 (3): 1-15. <http://dx.doi.org/10.4403/jlis.it-12653>.
- OpenAIRE. n.d. "Argos data management tool." Accessed 7 April 2021. <https://argos.openaire.eu/splash/>.
- Pradhan, Sanghamitra. 2018. *Cataloguing of Non-Print Resources: a Practical Manual*. New Delhi: Ess Publications.
- Rice, Robin, and John Southall. 2016. *The Data Librarian's Handbook*. London: Facet Publishing.
- Svantesson, Lotta, and Monica Steletti. 2019. "DSpace ORCID integration: name authority control solution at the European University Institute." Presented at the The 14th International Conference on Open Repositories (OR2019), Hamburg, Germany. <https://doi.org/10.5281/ZENODO.3553926>.
- University of Southampton. "Registry of Open Access Repositories." Accessed 7 April 2021. <http://roar.eprints.org/>.
- Wilkinson, Mark D., Michel Dumontier, Barend Mons et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18>.

Integrated Search System: evolving the authority files

Elena Ravelli^(a), Maria Cristina Mataloni^(b)

a) Istituto Centrale per il Catalogo Unico - ICCU, <http://orcid.org/0000-0001-6402-2039>

b) Istituto Centrale per il Catalogo Unico - ICCU, <http://orcid.org/0000-0002-2791-2822>

Contact: Elena Ravelli, elena.ravelli@beniculturali.it;

Maria Cristina Mataloni, mariacristina.mataloni@beniculturali.it

Received: 1 April 2021; **Accepted:** 21 May 2021; **First Published:** 15 January 2022

ABSTRACT

The coexistence of separate authority files within the main databases managed by the ICCU for entities of the same kind is going to be superseded in the new SRI portal through the integration, at the level of cooperative application, of the authority files for EDIT16 and Manus OnLine with those of SBN. The clustering of authority files, made possible through batch procedures and services provided by the applicative protocol SBNMARC, is intended to the development of browsable links between different representations of the same entity. The presence of identifiers and link keys between informative objects is therefore crucial to match data from the specialised databases EDIT16 and Manus OnLine, stored in the digital aggregator Internet Culturale, and shared through the collective catalogue SBN, with diverse quality and model but referred to the same resources and entities. The cluster of entities will be built upon the SBN Index, according to the quantity of data already available in its authority file and to exploit existing services and infrastructures which make shared cataloguing possible. SBN will also provide the spine of the integrated representation of entities through the public access platform of the new portal.

KEYWORDS

SRI; SBN; EDIT16; MOL; Alphabetica.

After the 1966 flood in Florence, which caused extensive damage to the library collections and catalogues of the Biblioteca Nazionale Centrale di Firenze, the Centro nazionale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, known since 1975 as the Istituto per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), decided to proceed with the microfilming of card and paper catalogues in State libraries. This was the first step towards the digitisation of historical catalogues of the library collections held at preservation libraries; digitisation would guarantee the very survival of the catalogues and furthermore would offer access to their contents to a very wide public, even from a remote location.

The initiative aimed at averting the risk of dispersion of the immense cultural heritage available in Italian libraries. This task, in which ICCU is still engaged today, is very demanding if we consider that the Anagrafe delle Biblioteche Italiane¹ registers approximately 12,000 libraries of different types, divided among state and local authorities libraries, university libraries, ecclesiastical libraries and cultural institution libraries. This is a snapshot, not yet exhaustive, of the Italian library reality, characterised by an extreme fragmentation from a geographical, organisational and institutional point of view, which is the manifestation of the historical and cultural events of the country.

This is the context in which the SBN (Servizio Bibliotecario Nazionale)², the network of Italian libraries, promoted by the Ministero dei beni e delle attività culturali e per il turismo in collaboration with regions and the university system, and coordinated by the ICCU, was born. It is based on an organisational model of cooperation and participation, designed to manage thousands of institutions. With this purpose in mind, in addition to the design and management of SBN and its database, the projects that led to the creation of ICCU's specialized bibliographic databases (EDIT16³ and Manus Online⁴) started in the 1980s. All these projects have been crucial in bringing libraries out of their isolation and pushing them to build a network to obtain a mutual advantage in terms of visibility. At the same time, moreover, the cooperative model has proved to be the only viable way to guarantee the technological infrastructures, the expertise and financial resources capable of supporting the profound changes and the great complexity that have affected the field of Library and Information Science over the last decades.

Now ICCU is called to take a further step forward, that is, to enhance and make available the enormous work carried out over the years by Italian libraries to the widest and most heterogeneous audience possible, with particular regard to digital resources that are becoming increasingly important in number and quality. This is how the project Integrated Research System (SRI) was born, which foresees the possibility of querying ICCU databases at the same time through the creation of a single access point for searching and returning results. A fundamental aspect in this regard is the integration project of the database authority files managed by the Institute.

¹ Anagrafe delle Biblioteche Italiane. Accessed June 3, 2021. <https://anagrafe.iccu.sbn.it/it/>

² Servizio Bibliotecario Nazionale - SBN. Accessed June 3, 2021. <https://www.iccu.sbn.it/it/SBN/>

³ Censimento delle edizioni italiane del XVI secolo - EDIT16. Accessed June 3, 2021. http://edit16.iccu.sbn.it/web_iccu/ihome.htm

⁴ Censimento dei manoscritti delle biblioteche italiane. Accessed June 3, 2021. <https://manus.iccu.sbn.it/>

Bibliographic databases of ICCU

Servizio Bibliotecario Nazionale – SBN

This infrastructural network is based on a stellar architecture whose centre is the Index (Indice), to which are connected the peripheral SBN Nodes (Poli), which include aggregations of libraries sharing resources, services, a user base and guidelines. The SBN Index, together with the management procedures, offers the services needed to establish the collaborative system that allows peripheral clients (the SBN Nodes) to share bibliographic information.

Currently there are nearly 6,600 libraries that have joined SBN, brought together in 104 Nodes. The size of the collective catalogue exceeds 18 million titles related to different types of material for a total of more than 102 million holdings.

Participation and cooperation are based on the sharing of common working methodologies, standards and uniform cataloguing rules as well as on a context characterised by flexibility in network participation. This flexibility allows Nodes to choose how to join on the basis of different profiles regarding the data to be shared with the collective catalogue. This context allows libraries to choose the quantity and quality of documents shared with the collective catalogue and the management of authority entries, and is essential in order to adhere to rules that make cooperation possible in a non-conflicting manner and consistent with the principles of the cataloguing involved.

On the catalogue front, ICCU has worked in recent years on the evolution of the SBN Index with the purpose of expanding the types of resources owned by libraries, including materials such as cartography, graphics, audio-visual, electronic resources and music. Moreover, in a highly modified context, both for the evolution of international standards and for the publication of the national cataloguing rules REICAT⁵, the realisation of the authority file has received particular care and attention. As of 2018, specific regulations have been developed for the registration of the different entities at the authority level in SBN. The quality of SBN catalogue data is a key value for user services. With this in mind, ICCU is working to strengthen and differentiate forms of cooperation through specific working groups, coordinated by ICCU, which involve librarians from different institutions, representative of the entire national territory, in the review and implementation of SBN authorities⁶.

EDIT16

The Censimento per le Edizioni Italiane del XVI secolo – EDIT16 was born in the 1980s with the purpose of documenting printed production from 1501 to 1600 in Italy and in the Italian language in other countries. Animated by a strong cooperative spirit, the project currently involves 1,597

⁵ *Regole italiane di catalogazione REICAT*. 2009. Roma: ICCU. <https://norme.iccu.sbn.it/index.php?title=Reicat>

⁶ The following working groups were launched in 2020: Working group for the management and maintenance of SBN Authority File. Printers, publishers, etc. ([https://www.iccu.sbn.it/it/attivita-servizi/gruppi-di-lavoro-e-commissioni/gruppo-di-lavoro-per-la-gestione-e-manutenzione-dellauthority-file-di-sbn-editori-tipografi-etc./](https://www.iccu.sbn.it/it/attivita-servizi/gruppi-di-lavoro-e-commissioni/gruppo-di-lavoro-per-la-gestione-e-manutenzione-dellauthority-file-di-sbn-editori-tipografi-etc/)) and Working Group for the management and implementation of the SBN Authority File. Names (<https://www.iccu.sbn.it/it/attivita-servizi/gruppi-di-lavoro-e-commissioni/gruppo-di-lavoro-per-la-gestione-e-manutenzione-dellauthority-file-di-sbn-nomi/>)

Italian and extraterritorial libraries. Since 2017, EDIT16 has expanded to record editions and holdings beyond national borders, confirming the transformation of the Census from a catalogue to a fundamental bibliographic resource for the study of Italian Renaissance culture. The first foreign institution to join EDIT16 was the British Library, which made a significant contribution to the database, thanks to the wealth of its collections of Italian editions.

In addition to the database reserved for bibliographic descriptions, the authority file, characterised by a highly specialised level of detail, has been developed including personal names, printing places, publishers/printers, and printer devices. EDIT16's authority file comprises structured data according to the descriptive standards of authority records with appropriately diversified elements to ensure their peculiarities. The quality and homogeneity of the data and the consistency of the information in the various databases have been ensured by the research work carried out by the editorial group, set up within the ICCU's Area di attività per la bibliografia, la catalogazione e il censimento del libro antico; since the beginning of the project, the group has taken on the management of the centralised collection of bibliographic descriptions, the definition of a working methodology, and the registration of authority records.

Two independent collateral databases, still closely related to the larger database, are the bibliography database, which describes the printed and electronic references listed in EDIT16, and the dedications database.

Digitisation is a crucial feature in EDIT16. More than half of the bibliographic records are complemented by title page and colophon images, and the description of each printer device includes its image. In recent years a considerable number of links to complete Italian and foreign digital copies available on the web has been added to EDIT16.

MANUS Online

Launched in 1988 with the coordination of the ICCU's Area di attività per la bibliografia, la catalogazione e il censimento dei manoscritti, Manus Online (MOL) is the first national project focused on the recognition and cataloguing of the immense manuscript heritage in the Latin alphabet from the Middle Ages to the contemporary age and preserved in Italian libraries. In recent years more and more space has been allocated for the census of papers (15th-20th centuries).

Since 2007, Manus Online has been available as an application arranged into specialised fields that include both a catalogue available to users and a cataloguing module, available free of charge for conservation organisations (public, private, ecclesiastical) participating in the project. Manus Online was the result of a collaboration among librarians, manuscript scholars and computer scientists who worked together to create a platform, which, all the while ensuring respect for the traditional descriptive elements of manuscripts, was more in line with international standards to allow an exchange of data and a dialogue with other bibliographic databases. Manus Online is based on an XML/TEI data schema, which was the most suitable format for exhaustively encoding the descriptive data of a manuscript and for allowing data to be exchanged with other data models as well, without loss of information.

There are 415 conservation and research organisations currently involved in the Manus Online project. In addition to librarians, individual scholars are also invited to propose variations to de-

scriptive data through the Forum, a section of the portal that allows for a constant exchange of opinions and suggestions with ICCU and libraries.

Of particular interest is the section “Special projects”: a module that allows the acquisition and management of specialist and international research projects. These projects, which maintain full organisational autonomy, use Manus Online as cataloguing software. The descriptions are offered to the public on each project’s own platform as well as on the Manus Online website in the section reserved for such projects.

In 2015 the Gruppo di lavoro per la gestione e la manutenzione dell’authority file of Manus Online was established with the purpose not only to monitor and correct the authority file in Manus Online but also to draw up guidelines with methodological procedures for the registration of authority records. The Manus Online section reserved for authority work is aimed at managing personal, collective, family and place names in printed sources and catalogued manuscripts.

Critical issues

As expressions of different projects, ICCU databases and information systems have been developed and managed over the years by separate offices. In the absence of central coordination to ensure an integrated development, these projects have been carried out using different softwares, languages and developmental approaches, thereby producing platforms based on very different data models and also data quality that meet the different expectations of their intended user communities.

Reference standards, even when held in common, have sometimes been adapted to the need to ensure the specificity of different objectives; moreover, in recent years, not all platforms have updated to the latest standards. As a result, at present, the same resource or entity can be recorded in different forms on individual platforms. A very clear example is the comparison between EDIT16 and SBN: as many users will have had the opportunity to verify, there is not always a biunivocal correspondence between the same resource described in the two databases. Just think of the case of different issues by date: in EDIT16 they are recorded in separate records whereas in SBN they have been described for many years⁷ as variants within the same record. On the contrary, the same EDIT16 identifier can be associated with several SBN identifiers if in SBN a record has been created for each volume in a multi-level description, as well as for the main record, whereas only one record describing the whole edition has been created in EDIT16.

Even in cases where the name is recorded in the same form in the databases, their data model changes considerably. If we analyse, for example, the authority record of a printer including the same basic information, the structure of the data varies considerably. In EDIT16, in fact, the printers authority file, particularly significant for this resource, is organised into several fields and links, some of which are in SBN.

In order to access the resources of the individual ICCU databases, which are described separately in the different environments, the user needs to start from the specific search interfaces available,

⁷ The more recent 2016 guidelines for early printed books in SBN, which have amended previous regulations, require that variants by date of the same edition should be described in different records. Still many corrections remain to be made on previous records.

in the absence of any connection among the platforms. The only exception, since 2016, is the introduction of the link between the bibliographic records of SBN and the corresponding records in specialised databases related to early printed books, including EDIT16. However, this link is not always available, being the result of individual cataloguing work.

Moreover, the specialised databases EDIT16 and Manus Online, and the SBN system allow for the management of digital attachments to descriptive records, but the current system of indexing and using digital resources (Internet Culturale and its index based on the BIB-MAG profile) does not recognise them. The latter makes integrated management of digital resources inefficient and complicated.

Internet Culturale: a partial solution

An integral part of the BDI (Italian Digital Library) project is *Internet Culturale, cataloghi e collezioni digitali delle biblioteche italiane*⁸, a portal launched by ICCU in 2005 to promote knowledge of the Italian book heritage through access to both catalogues and digital collections. It also offers cultural insights through multimedia resources (itineraries, exhibitions, authors and works, 3D-paths), dedicated to literary, scientific, artistic and musical culture.

What you get from a search in Internet Culturale, in addition to the Digital Index, is a set of results returned from each database queried: EDIT16, Manus Online, SBN, Historical Catalogues⁹ and Digital Library, the digital repository made available by ICCU for the Italian libraries. This is achieved by mapping data from ICCU databases on a common minimum Dublin Core profile. Search options are therefore limited, compared to what is offered by the different front-ends.

In view of the logical separation between records managed by different platforms, however, the Meta-Index system queries these databases, limiting itself to juxtaposing the information objects coming from the autonomous management environments. In addition, there are no engineered import procedures that ensure synchronisation in the alignment of Internet Culturale Indexes with those of SBN and of the specialised databases, thereby generating different search results by accessing the specific search sites.

Digital collections make up the Digital library, with more than 15 million associated digital files. The search in these databases is done through the Meta-Index of Internet Culturale by extracting data from the original databases. The data, whose characteristics, content and format vary, have been partially made consistent through a common profile based on the Dublin Core standard and qualified with the necessary extensions. However, the data profile in Internet Culturale is less rich than that of the same resources as they are recorded in SBN and the other specialist databases, and therefore does not fully represent and integrate the digital resources described through them. In addition, the data-feed process of the Digital Library of Internet Culturale is complex, rigid and rather expensive, especially for the institutions which do not have their own digital heritage management system and suppliers to provide this service.

⁸ Internet Culturale. Accessed June 3, 2021. <https://www.internetculturale.it/>

⁹ Cataloghi storici digitalizzati. Accessed June 3, 2021. <http://cataloghistorigi.bdi.sbn.it/>

Towards an integration of national bibliographic services: Integrated Research

In addition to the issues outlined above, there were cuts in professional resources, and, at the same time, the need to make new investments for the maintenance, updating and development of the individual platforms. As a result, we found ourselves having to rethink the structure of the information systems managed by ICCU in the twofold perspective of rationalisation and optimisation of resources, on the one hand, and of a new model of integration in data search and retrieval, on the other.

At the end of a process of reflection and in-depth analysis carried out within the framework of working groups promoted by the Direzione generale biblioteche e istituti culturali of the Ministero dei beni e delle attività culturali e per il turismo, the SRI project (Integrated Research System) was realised¹⁰. This project focuses on the integration of the databases managed by ICCU and at the same time will offer more effective services to meet the information needs of different user groups, ranging over from professional researchers up to the merely curious users. The project means therefore to overcome the fragmentation of the databases and rationalise the communication model of ICCU platforms, by creating a distributed information architecture that makes it possible to use the resources in the systems described above, including those in digital format, through a single access point.

The solution adopted to maintain information consistency among the elements of EDIT16, Manus Online and Internet Culturale Digital Index is achieved by configuring these systems in Client server mode with the SBN Index acting as a joint element and guide between these different systems. The new EDIT16 and Manus Online are, in fact, configured as specialist SBN nodes in the new set-up and are able to contribute to the collective catalogue while maintaining their own specificity and autonomy. The aim is to share a large part of the information with the reference record in the Index, with the future goal of sharing even greater information, once the system has been consolidated.

With this in mind, the back-end applications of EDIT16 and Manus Online have been re-engineered to enable dialogue with the central Index; at the same time, the Internet Culturale system has been optimised.

In order to achieve this goal, integration of the databases must be foreseen as early as in the creation phase of bibliographic records (only for EDIT16) and of the most important elements of the bibliographic data, such as names. The structural coherence between the databases is ensured by joint keys (mutual references) among the different information objects that refer to the same entities. This solution allows SRI to recognise that two representations refer to the same object. In particular, a resource X or an entity Y will be retrieved only once through the single search point, whereas specific representations, which are diverse for data models, for quality, or purpose, will be reached through links. Links will indeed allow navigation through the search interfaces, which will maintain their own functions and specificities designed for their own user community. As to the Manus Online database, the new re-engineered management application provides for the implementation of an exchange module exchange module, which allows sharing only authority

¹⁰ For an overview of the steps that led to the development of the Integrated Research System, see Patrizia Martini. 2018. "Verso un'integrazione dei servizi bibliografici nazionali." *DigiItalia* no. 2 (2018): 9-16.

records (including both intellectual responsibilities for texts and material responsibilities related to codicological features) with the SBN Index and which are aligned through specific services of the SBNMARC protocol, the dialogue protocol allowing the exchange of data between the central index and the SBN nodes. Any additional fields needed to complete the information in the Manus Online authority record are stored locally and not shared with the SBN Index.

Internet Culturale will continue to serve as an infrastructure dedicated to the collection and indexing of digital copies in its own repository and to those made available by remote repositories, but it will no longer have its own website, including a search engine and other tools available to users. In the architecture of the new information system, it will provide all those digital resources which can be linked to a cataloguing record in the main database to the central integration and indexing system, while also enriching the bibliographic core with all the technical and usage details in order to enhance its research tool on digital heritage.

Moreover, to widen the range of external providers of digital content, the aggregator will only allow the acquisition of descriptive metadata, leaving to the digital repository the function of making its digital content publicly available to users.

The process of authority files integration

As for bibliographic records coming from the SBN Index and the EDIT16 database, SRI has planned the unification of authority records. In particular, we refer to the files of personal and collective names of SBN, EDIT16 and Manus Online, and also, as far as the EDIT16 and SBN databases are concerned, to the printers devices and printing places.

Whereas inconsistencies between bibliographic records can be merely the result of cataloguing choices (e.g., the option of multilevel cataloguing), inconsistencies between names are rather substantial, and may prevent the ability of end users to identify an object to be the same in SBN, EDIT16 and Manus Online.

For instance, the authority records for Saint Roberto Bellarmino occurs in the three databases in three different forms:

1. Bellarmino, Roberto in SBN
2. Roberto : Bellarmino<santo> in EDIT16
3. Bellarmino, Roberto <santo ; 1542-1621> in Manus Online

The crucial role of a reconciliation among the three authority files is therefore clear, not only for end users but for cataloguers as well, who will be able to cross-reference authority records.

This process of integration of the authority files will be performed through a manual and painstaking double check of the databases. Some automatic procedures specifically developed will be useful at this stage to highlight problematic instances which require human intervention.

These procedures, which will concern the EDIT16, Manus Online and SBN authority files, will be addressed to catch duplicates or clearly erroneous forms. Such issues are rather frequent in SBN and Manus Online, whose data are the result of shared cataloguing, more exposed according to its nature to the risk of inconsistencies. This is not much of an issue for EDIT16, as its authority files are the result of research work carried out by the ICCU at a central level. However, EDIT16 records are often inconsistent with those of SBN from a formal point of view; indeed, EDIT16

authority records have been created according to the previous Italian cataloguing rules for authors (RICA¹¹) standards, as the database has been developed before the issue of the REICAT code in 2009. There has been no chance to update the authority records to the new cataloguing rules to date.

After these preparatory cleaning tasks, we will go ahead with the identification of SBN records matching records in specialised databases. In this process, SBN IDs will be added to the link field of the specialised databases, as well as the identifiers of the specialised databases to the SBN database. This will ensure data persistency, and their mutual reference will also be guaranteed. The inclusion of these relationships within the UNIMARC exchange files, meant to feed the new search interfaces, will make possible the aggregation of clusters of information objects by SRI's indexing engine.

The joint keys, which are built automatically within previous databases, will still be created when the new system is fully operational through the cooperation procedures. For this purpose, the matching algorithms, defined and refined within the development of the procedure described above and named 'Import-as-recognition', will also be included in the re-engineered specialised applications. The term 'import' means that names (of personal and collective entities) only included in EDIT16 and Manus Online will also be added to the Index database, which will store the whole set of authority records.

During the import procedure, the three authority files of SBN, Manus Online and EDIT16 will continue to be enriched and amended by the work carried out at the central level, as well as at a peripheral level by Nodes and libraries, both in the Index and in the specialised databases (as in cataloguing, data correction, etc.). In order to avoid the risk of misalignment and of jeopardising the consistency of the work done, the EDIT16 and Manus Online management systems will be re-engineered before performing the clustering procedures.

The reference information sheet, represented by the SBN record, is thus enriched with links to specialised databases, which, in turn, will have the chance to reach a larger audience, having their results listed not only in the new portal but also in the new OPAC SBN. Manus Online and EDIT16 will still maintain their own representation of entities in the reference systems as a result of the expertise of each database.

If these activities and tools prove to be apt to the task, similar procedures will be undertaken for other authority files, that is the printer devices database in EDIT16 and place names ones in both EDIT16 and Manus Online.

Integrated Research System – SBNTeca – Main services

As previously mentioned, the integration of the entire ecosystem also involves substantial correlation with digital resources and related metadata, stored and managed by the Digital Library system.

Therefore, an organisation of the complex system of aggregation and fruition of the digital resources of ICCU and Internet Culturale is in progress, as a companion to the development of the

¹¹ *Regole italiane di catalogazione per autore*. 1982. Roma: ICCU.

new management applications. Basically, the new architecture of the digital flow allows digitised items to be in fact an extension of the catalogue through the link between digital copies and the bibliographic records, which was often missing in the information set provided in the previous systems.

Another fundamental component of the SRI is therefore the SBNTECA, which is a digital library capable of allowing the management of digital objects (images, audio-visual documents, etc.) within individual SBN Nodes and their exposure to the central SRI system and, from here, their display through the central SRI system as well.

SBNTECA, besides allowing the management of digital objects, is also used for the creation, import and management of metadata associated with such digital objects (technical metadata), as well as aggregates between them (e.g., the pages of one book). To do this, it must be able to act on the main metadata standards for digital content: MAG, a standard for management and administrative metadata, and METS (mainly in the Google-METS and METS-ICCU specifications).

The services made available by SBNTECA are also meant to recover the ‘submerged digital’, often included in often difficult to access share due to preservation contexts, as well as poorly valued or only available in off-line environments managed by Italian libraries. Also, the new digitisation campaigns find, in this context, an efficient system of management and fruition of digital and multimedia content.

Network of portals¹²

The multi-layer work described so far, mainly addressed to reconcile authority files, proves to be meaningful for end users in the results of the new platforms. Each database presents indeed a new faceted search interface providing features made possible by the new architecture. It is, in fact, a network of portals, each of which gives access to research services and types of content intended for different communities of users characterised by diversified information needs.

The case of Manus Online is significant under this respect, as it offers a more articulated internal representation of the manuscript record and a clearer identification of the textual units and their grouping in codicological units.

Functionalities will not be weakened; on the contrary, the system as a whole will provide new search options in a more technologically advanced and optimised context.

SBNTECA’s services, which are integrated into the management systems of specialised databases, make direct representation of digital content available through the ecosystem’s central viewer, Mirador, based on the IIIF protocol.

Alongside the re-engineering and functional review of the specialised research platforms, the project has planned a similar intervention on the portal of the SBN catalogue. The main difference is that the SBN OPAC will provide an integrated search within the bibliographic records of EDIT16 and Manus Online. The SBN catalogue will be complemented by links to records coming from EDIT16 and Manus Online; EDIT16 records can either be referred to resources also includ-

¹² Cerullo, Luigi, and Maria Cristina Mataloni. 2020. “Sistema di ricerca integrato: un nuovo catalogo di servizi per le biblioteche.” *DigitItalia* no. 2 (2020): 16-25.

ed in SBN as part of the collections of one or more SBN libraries, or to resources not included in SBN when part of collections of libraries not participating in SBN. Manuscripts coming from Manus Online, on the other hand, are resources of a type not managed in the collective catalogue to date. In this case, internal descriptions of manuscripts are anyway retrieved, i.e. records including bibliographic elements and authorities, whose descriptive profile is better suited to the data return model of the shared catalogue.

ALPHABETICA: a new portal for Italian libraries¹³

To the eyes of end users, the most relevant new tool will be the new portal of Italian libraries, Alphabetica. The bibliographic core, represented by the general catalogue and its logic integration, is enriched by the data coming from other related databases managed by the ICCU, such as the portal *1418 – documenti e immagini della grande guerra*¹⁴, the historical catalogues, virtual exhibitions built with *Movio– Mostre virtuali Online*¹⁵ (Digital Online exhibitions). Moreover, the architecture of Alphabetica will possibly allow the integration of more databases potentially interested in joining the Alphabetica network. The logic behind the portal goes beyond the traditional model of bibliographic research and restitution of an OPAC, even of an advanced one, and is an attempt to build a proper search model based on a solid but flexible system of taxonomies.

Alphabetica classifies all kind of resources and related entities (names of collective entities, places, persons) according to a dual classification system to comply both with the traditional classification of material in SBN, and with controlled vocabularies for the classification of objects in order to arrange them within thematic channels. This approach will provide a way round to the oddity of semantic terms in the catalogue, which we are in the meantime trying to address through rather complex off-line procedures on the SBN Index.

The stages required for the analysis, planning and testing of Alphabetica are obviously along and painstaking process in order to realise an innovative and effective new reference, which will be able to exploit and showcase the longstanding and valuable work carried on in Italian libraries, not only for a specialist audience but open to a new and diverse user base.

¹³ Buttò, Simonetta. 2020. "Alphabetica, il nuovo portale per la ricerca integrata: un salto di qualità per le biblioteche italiane." *DigItalia* no. 2 (2020): 9-15.

¹⁴ <http://www.14-18.it/home>

¹⁵ <https://www.movio.beniculturali.it/>

References

- 1418 Documenti e immagini della grande guerra*. Accessed June 3, 2021. <http://www.14-18.it/home>.
- Anagrafe delle Biblioteche Italiane*. Accessed June 3, 2021. <https://anagrafe.iccu.sbn.it/it/>.
- Buttò, Simonetta. 2020. "Alphabetic, il nuovo portale per la ricerca integrata: un salto di qualità per le biblioteche italiane." *DigItalia*, no. 2 (2020): 9-15. <http://digitalia.sbn.it/article/view/2624>
- Cataloghi storici digitalizzati*. Accessed June 3, 2021. <http://cataloghistorici.bdi.sbn.it/>.
- Censimento delle edizioni italiane del XVI secolo - EDIT16*. Accessed June 3, 2021. http://edit16.iccu.sbn.it/web_iccu/ihome.htm.
- Censimento dei manoscritti delle biblioteche italiane*. Accessed June 3, 2021. <https://manus.iccu.sbn.it/>.
- Cerullo, Luigi, and Maria Cristina Mataloni. 2020. "Sistema di ricerca integrato: un nuovo catalogo di servizi per le biblioteche." *DigItalia*, no. 2 (2020): 16-25. <http://digitalia.sbn.it/article/view/2625>
- Internet Culturale*. Accessed June 3, 2021. <https://www.internetculturale.it/>.
- Martini, Patrizia. 2018. "Verso un'integrazione dei servizi bibliografici nazionali." *DigItalia*, no. 2 (2018): 9-16. <http://digitalia.sbn.it/article/view/2162>.
- MOVIO Mostre Virtuali Online*. Accessed June 3, 2021. <https://www.movio.beniculturali.it/>.
- Regole italiane di catalogazione per autore*. 1982. Roma: ICCU.
- Regole italiane di catalogazione REICAT*. 2009. Roma: ICCU. <https://norme.iccu.sbn.it/index.php?title=Reicat>
- Servizio Bibliotecario Nazionale - SBN*. Accessed June 3, 2021. <https://www.iccu.sbn.it/it/SBN/>.

DREAM. A project about non-Latin script data

Antonella Fallerini^(a), Agnese Galeffi^(b), Andrea Ribichini^(c),
Mario Santanché^(d), Mattia Vallania^(e)

a) Sapienza Università di Roma

b) Sapienza Università di Roma, <https://orcid.org/0000-0003-0799-5699>

c) Sapienza Università di Roma, <https://orcid.org/0000-0002-0281-4257>

d) Sapienza Università di Roma, <https://orcid.org/0000-0003-1777-1162>

e) Sapienza Università di Roma

Contact: Antonella Fallerini, antonella.fallerini@uniroma1.it; Agnese Galeffi, agnese.galeffi@uniroma1.it;
Andrea Ribichini, ribichini@diag.uniroma1.it; Mario Santanché, mario.santanche@uniroma1.it;
Mattia Vallania, mattia.vallania@uniroma1.it

Received: 5 May 2021; **Accepted:** 16 June 2021; **First Published:** 15 January 2022

ABSTRACT

The DREAM project is a large research project funded by Sapienza University of Rome, dealing with bibliographic data in non-Latin scripts. As the National Bibliographic Service catalogue (SBN) does not yet manage data in non-Latin scripts, the aim of DREAM is to offer researchers a catalogue searchable through original scripts (such as Arabic, Chinese, Cyrillic, etc.). One of the most remarkable features of the project is the creation of an ILS-independent working context in which the catalogue may find and retrieve data in original script from authoritative catalogues, starting from the existing romanized ones. From a technical standpoint, the ever increasing Unicode support offered by modern operating systems, DBMSs and indexing engines makes the rapid development of the relevant software tools a concrete possibility. This in turn implies a shift in scientific focus towards the (often subtle) record linkage operations between different data sources. The authors hope that the DREAM project will gather the adhesion of other Italian libraries that perceive the same needs. Furthermore, as soon as SBN will support the management of data in non-Latin scripts, the DREAM project partners will be able to contribute with their data.

KEYWORDS

Romanization; MARC records; Cataloguing; Transliteration.

Non-Latin script cataloguing. The context

The DREAM (Data Recording Entry Alternative Multi-script) project was born in Sapienza university in 2019 in order to create a repository for bibliographic data in non-Latin scripts, publicly available as a cooperative catalogue. This need arises from the evidence that the SBN national catalogue, to which Sapienza libraries adhere as do thousands of other Italian libraries, does not fully support the UTF-8 character encoding. SBN catalogue is based on shared cataloguing: all the participant libraries contribute sending data from the local nodes (that is, aggregation of libraries) to the central index. Member libraries may use a variety of LMS authorized by the ICCU, the national agency in charge with the SBN catalogue maintenance, but regardless of the software capabilities, the central index accepts data in Latin script only. All the languages expressed through other scripts, such as Arabic, Chinese, Cyrillic, Hebrew, Japanese, Greek, must be transliterated using the ISO instructions. This requirement is stated in the Italian cataloguing rules, REICAT (ICCU 2016b), and restated by the SBN cataloguing instructions (ICCU 2016a). The transliteration process has many disadvantages for both the involved actors – cataloguers and users.

Notwithstanding some attempts at automatic transliteration (Eryani 2021) and the availability of online tools (DuBose 2019), this activity is a very time consuming one presenting a large number of technical problems (Ismail and Md. Roni 2010), not to mention the variety of connected cataloguing issues such as

- The use, in some contexts, of unsound transliteration scheme (Molani 2006).
- The transliteration and conversion of personal names, place names, corporate bodies, and other entities (Li 2004).
- The subject access (El-Sherbini and Chen 2011).

Besides that, the equity of access – one of the basis of the ethics in library science – is not guaranteed since those who need these materials have to determine how cataloguers could have transliterated the data. On the opposite, the users who want roman script resources are not required to make this extra and inefficient effort (Agenbrood 2006, 22). For users who are native language speakers in non-Latin scripts, the transliteration is totally useless since they have all the knowledge and tools to perform a search in the library catalogues using the original script.

For all other users reading in second language, transliterated text requires an additional cognitive demand since they typically acquire and access a cohesive set of phonological, orthographic (and possibly semantic) representations of words in their second language, whereas transliteration requires readers to create cross-script associations between phonological-semantic representations in one language and previously unrelated orthographic forms in another (Rao, Mathur, and Singh 2013, 205). All these elements cause many obstacles in users' searching ability and accessibility in retrieving bibliographic records (Kim 2006)

The DREAM project

The DREAM project is a major university project funded by Sapienza University of Rome in 2019; the project leader is Federico Masini, professor of Chinese at the Istituto Italiano di studi orientali (ISO) of Sapienza university of Rome and the research staff is a mix of academics and library per-

sonnel, both involved on the front line of the activities. The very first idea of the DREAM project arose in May 2018, when Antonella Fallerini, librarian at the ISO library, joined the workshop “Building a Network of Korean Resources Specialists in Europe”, organized by Freie Universität Berlin – Campus Library and funded by the Korea Foundation. The workshop aimed at bringing together European Korean Studies librarians in order to develop a professional network within Europe and strengthen the representation of interests of Korean Studies librarians in national and worldwide library information structures. While discussing with the colleagues attending the workshop, Fallerini highlighted the severe limitations of transliterated data compared to bibliographic descriptions in original scripts. The expected result of that was that some colleagues confirmed they did never find a single record in original script in Italian catalogues. Their obvious self-explanation was that there was a great scarcity of our collections in Far Eastern languages. The available online catalogues do not give any advice about the transliteration and researchers have no reason to expect such a treatment. The extensive transliteration practice in cataloguing has as a direct consequence the underrepresentation of our library collections both at national and international level. To give an idea of the extend of the phenomenon, an internal preliminary investigation conducted on the online catalogues of the most representative Italian institutions, such as larger universities and research libraries, have shown that more than 500,000 resources have been catalogued in romanization. It is possible to estimate that there are at least double that number waiting to be catalogued.

What is the aim of DREAM?

DREAM project aims to figure out a provisional and cooperative solution in order to create a repository for non-Latin scripts data, available as a catalogue in the near future. At the present moment, the cataloguers must create transliterated data to feed SBN; if adhering to the DREAM project, the cataloguer will also use DREAM tools to search for the corresponding record in original script in authoritative catalogues. Once the possible matches have been identified, the system will present them to the cataloguer, giving him/her the responsibility of confirming or dismissing them. These data will complement the transliterated ones that are already being produced for SBN shared cataloguing.

The result of this procedure will be a cluster of records that will be show in the DREAM catalogue giving the user the possibility to make searches using the preferred forms (in original scripts or in transliteration, even according to different schemas).

The DREAM project do not want to propose an SBN-alternative context. On the opposite, its features are developed taking into consideration both the respect of SBN rules and its potential developments. When SBN will accept data in non-Latin scripts, the libraries adhering to DREAM will have the possibility to feed their records into the national catalogue. This is why the DREAM project has a provisional nature. The adjective “provisional” may be referred to two aspects: first of all, it connotes the research aspect. DREAM’s aim is to produce a working solution and at the same time, to explore, to verify, and to find the best ways to achieve the projects’ goals. Since Sapienza libraries are part of the SBN network, there is no intention to create a new network or some alternative solutions; this is the second significance of provisional. DREAM project wants to create an environment where the cataloguer can retrieve data in

non-Latin scripts, match them with the available transliterated ones and make them available in a specific DREAM catalogue. Both the working environment and the user search interface will be independent from SBN as well as from the software used by the libraries joining the project. We hope in fact that other Italian libraries – even those not members of SBN – will be interested in joining the DREAM catalogue, once some of the fundamental components of the its architecture have been realized. The DREAM project is still ongoing. We would like to stress one of the project’s strengths: flexibility. We have a clear idea of the final results we want to achieve, but there is no prejudice about how to reach them. There are just some constrains due to the cataloguing context we have to dialogue with at some point, that is the software used and the SBN catalogue.

Main points

DREAM is ILS independent

Commercial software available to librarians are built to maximise output (cataloguing, lending, library management, etc.) and are therefore, in most cases, designed to be stable and standard. If you need a flexible environment to, for instance, carry out an experiment or a research project, it is difficult to balance these development needs – maybe even unsuccessfully – with the commercial logic of software distributors. Anyway, Sapienza has invested in our ILS (SebinaNext) in order to implement in the near future some new features, such as to accept, manage and visualize data in non-Latin script and especially right-to-left scripts, to handle VIAF id and an OAI-PMH module for authority data. DREAM will be an external and ILS independent environment. This need would not have arisen if we had a flexible and welcoming open source software or library platform in use. In this case, the DREAM project would have been just another component, a small one, of a larger system. What we learnt: often the paths you thought you were taking do not turn out to be fruitful and you have to go back, change your path and sometimes even rewrite the map. These features of research projects do not match the market logic.

Retrieve bibliographic data from reliable sources

In order to quickly populate the DREAM catalogue, we are going to start from the traditional transliterated records already existing, to search for equivalent records in original script in authoritative catalogues, and to import them. These procedures present certain degrees of difficulty. First of all, the identification of reliable sources to retrieve data. This is a scientific but also a technical task. It is not only a matter of knowing the most representative institutions for the languages of interest, but also of selecting those that have a data format easy to manage or map and an accessible retrieval option.

The current DREAM implementation supports this “search and match” between, on one side Sapienza University of Rome catalogue and on the other side the Bibliothèque nationale de France, the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3kat), and the Système universitaire de documentation (SUDOC). Since these sources expose their data through a variety of protocols, such as OAI-PMH, SRU and Z39.50, different clients are needed. More-

over, to process data, specific response parsers are required for each source. As a matter of fact, even though the retrieved data are in standard formats (MARC21, UNIMARC), the packaging of data varies from source to source, containing error messages and paging information in ad hoc formats. Even in environments that we assume to be highly standardized (dealing with MARC, Z39.50, SRU, OAI-PMH formats) we found, in addition to the expected MARC21-UNIMARC dichotomy, USMARC or local dialects of MARC, Dublin Core, and several application profiles. In order to obtain a presumed match of the data, different analyses and mappings are required each time for their retrieval and processing.

Different sources (we are talking about national bibliographies/catalogues and national library catalogues) also have different approaches to standards.

For example, MARC21 allows to put in the same record data in the original script (e.g. Cyrillic) together with transliterated data by using the combination 880 and \$6 but the cataloguing agency can choose whether to put in 880 the original script or the transliterated version. This allows the creation of (at least) two versions of the record. Moreover, the different granularity of the data contributes to make the match uncertain.

Authority data

Obviously, within the DREAM environment, in addition to bibliographic data, it is essential to import, manage and use authority data. In this respect, VIAF is the point of reference. Since the VIAF id is widely used, it is not only possible to retrieve authority clusters, but also to use the VIAF id as a bridge to navigate through catalogues in search of other bibliographic data of potential interest.

What we are building. The DREAM architecture

We designed a flexible, modular and scalable software architecture for a multiscript MetaOPAC (see Figure 1), based on the data warehousing paradigm (Inmon 2005; Kimball et al. 2008). We also developed a prototype implementation for research purposes (i.e., feasibility assessment, experimental evaluation of adopted solutions). The following is the description of our architecture's main components.

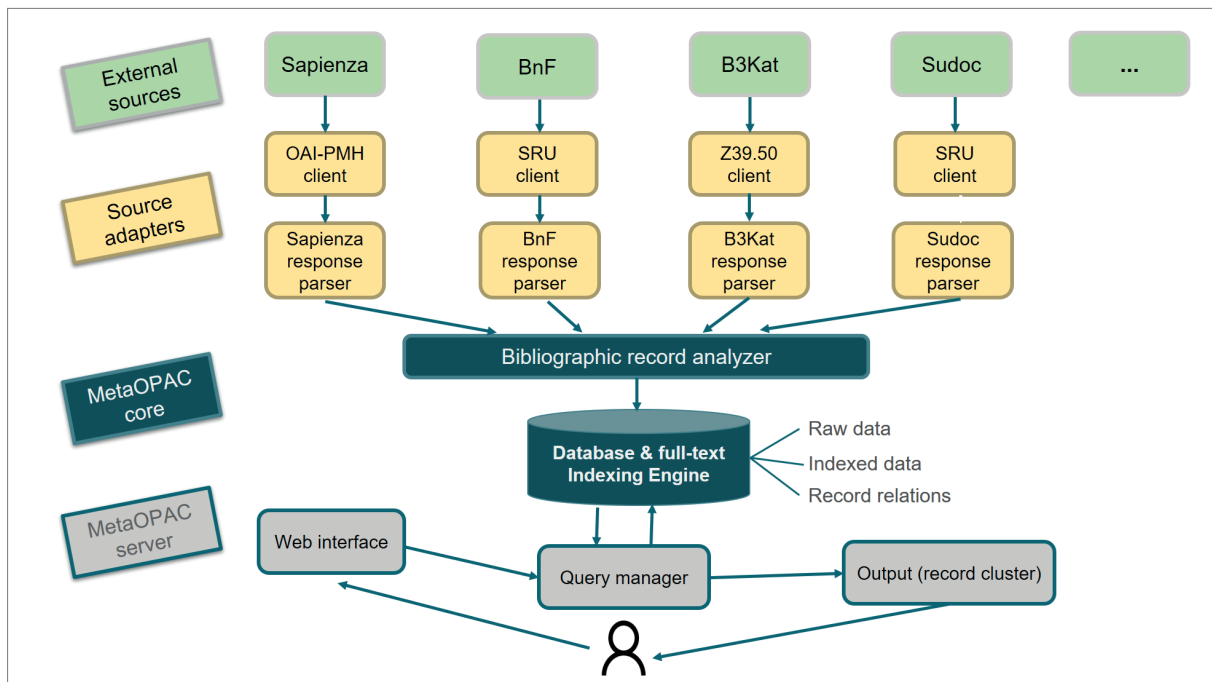


Fig. 1. MetaOPAC architecture

Source Adapters. We have taken into consideration and tested several data sources. The current implementation supports Sapienza University of Rome’s own catalogue, the Bibliothèque nationale de France (BNF), the Library Union Catalogue of Bavaria, Berlin and Brandenburg (B3Kat), and the Système universitaire de documentation (SUDOC). These sources make their data available through a variety of protocols, such as OAI-PMH, SRU and Z39.50. Therefore, clients are needed for each of these protocols. Moreover, an ad hoc response parser is required for each data source. This is because, even though the returned data are provided in standardized formats (e.g., UNIMARC, MARC 21), the packaging of these formats varies from source to source.

MetaOPAC Core. Downloaded bibliographic records are stored in a (relational) database, analyzed, and portions that are relevant for future search queries, e.g., title, authors, publisher (including all variants in both native script and transliteration, if present) are saved separately and properly indexed. Our prototype implementation currently uses MySQL as DBMS. The database structure consists of three tables:

- Table “raw” contains the unprocessed downloaded records.
- Table “indexed_data” contains, for each record, the extracted data to be indexed in order to speed up searches. At the present moment, we rely on MySQL’s full-text indexing capabilities (a recent addition). We remark that different scripts require different indexing methods: alphabetic and syllabic scripts are handled by the default token-based full-text indexer, with minimum token size set to 1 and stop words exclusion disabled, while Ideographic scripts are instead dealt with by an n-gram based indexer, with n=2.
- The third database table, “relations” represents associations between records from different data sources, that we call “clusters”. Clusters may be established through several methods (that we will discuss shortly).

MetaOPAC Server. Searches in our prototypal MetaOPAC implementation can be run through a web server that accepts HTTP GET requests. In addition to the traditional search criteria (keywords, title, author, publisher), wildcards and boolean operators are accepted. A query manager translates the searches into full-text database queries. The search results are returned as an XML document listing retrieved clusters sorted by *relevance* (a measure of the adherence of the records in each cluster to the search criteria).

How to feed the DREAM. Record linkage among data sources

In our MetaOPAC application, the construction of clusters (i.e., groups of records referring to the same entity) may be carried out through three methods.

1. *Manual Intervention.* The cataloguer manually identifies the correspondences between records from different data sources. In our prototype we have created 27 Sapienza-BNF pairs, 27 Sapienza-B3KAT pairs, and 40 Sapienza-SUDOC pairs. It is hoped that, as the number of partners grows, more and more librarians will contribute their associations across data sources to the MetaOPAC database.
2. *Identification by Unique Identifiers.* A second way to identify correspondences between records from different data sources is through unique identifiers. In our prototype we have used ISBN to search for matches (all supported external catalogues allow ISBN-based searches through their APIs).

Document Language	Sapienza Records with ISBN	Sapienza-BNF ISBN-based Matches	Sapienza-B3Kat ISBN-based Matches	Sapienza-SUDOC ISBN-based Matches
ARA	369	122 (33.06%)	113 (30.62%)	126 (34.15%)
CHI	1875	98 (5.23%)	492 (26.24%)	399 (21.28%)
HIN	25	8 (32%)	3 (12%)	8 (32%)
JPN	1771	246 (13.89%)	692 (39.07%)	781 (44.10%)
KOR	2191	80 (3.65%)	432 (19.72%)	457 (20.86%)
PER	66	7 (10.61%)	12 (18.18%)	15 (22.73%)
SAN	73	17 (23.29%)	26 (35.62%)	28 (38.36%)
SWA	1	1 (100%)	1 (100%)	1 (100%)

Table 1. Breakdown of positive search results

3. *Algorithmic Techniques.* The third method consists of a blend of well-established record linkage algorithmic techniques and ad hoc solutions. We proposed the following workflow, based on the VIAF:
 - Given as input a bibliographic record, we extract the VIAF code of its author (assumed to be present).
 - We then run a search on the VIAF online service for the extracted id, obtaining the variant form of the author's name used by each data source.

- For each supported source, a search-by-author, using the variant form obtained through VIAF as input string, is performed. This allows us to restrict the search domain to the works of that author.
- Finally, we run any record linkage algorithm we see fit in order to identify the correct matches between the input record and the records retrieved from the other data sources.

Standard record linkage techniques include the use of string similarity measures (Navarro, 2001) – Levenshtein distance (Levenshtein, 1966) being a popular one – to assess correspondences between fields such as title, subtitle and publisher (including their variants and versions in original script, if present). Comparison of other metadata (e.g., publication dates) may also be useful as a verification tool. Moreover, if the bibliographic record belongs to a cluster in the MetaOPAC database, then all metadata of the cluster may be used to identify the correct match. More sophisticated, domain-specific techniques may include transformations from one transliteration standard to another, and switching from original script to transliteration and vice versa. Early testing on 19 Sapienza records, manually matched with both BNF and SUDOC to provide a “ground truth”, has shown correct results in 17 cases. This is quite promising considering that for this test only the minimum normalized Levenshtein distance (i.e., Levenshtein distance divided by the length of the longest input string) between all title variants has been considered as a criterion.

Further steps

The project next steps are:

- Engaging partner institutions: we hope that this conference will also be an opportunity to promote the project and involve other partners who share the problem with data in non-Latin scripts
- From a technical standpoint, further tasks would include writing adapters to support additional sources, and launching larger scale algorithmic record linkage runs with feedback loops involving manual sample validation and fine-tuning of algorithmic features. Identified clusters should then be fed into the MetaOPAC prototype implementations, with measurement of both load and query times, in order to determine performance-critical sections that may need refinement both at the implementational and the architectural level.
- It is also needed to develop all the interfaces, both the back office minimal interface to allow cataloguers to validate the matches between records and the public DREAM catalogue search interface.

Bibliography

- Agenbroad, James E. 2006. "Romanization Is Not Enough." *Cataloging & Classification Quarterly* 42 (2): 21-34. https://doi.org/10.1300/J104v42n02_03
- DuBose, Joy. 2019. "Russian, Japanese, and Latin Oh My! Using Technology to Catalog Non-English Language Titles." *Cataloging & Classification Quarterly* 57 (7-8): 496-506. <https://doi.org/10.1080/01639374.2019.1671929>
- El-Sherbini, Magda, and Sherab Chen. 2011. "An Assessment of the Need to Provide Non-Roman Subject Access to the Library Online Catalog." *Cataloging & Classification Quarterly* 49 (6): 457-483. <https://doi.org/10.1080/01639374.2011.603108>
- Eryani, Fadhl, and Nizar Habash. 2021. "Automatic Romanization of Arabic Bibliographic Records." <https://arxiv.org/pdf/2103.07199.pdf>
- ICCU. 2016a. "Guida alla catalogazione in SBN – Materiale moderno." Last modified July 13, 2016. https://norme.iccu.sbn.it/index.php?title=Guida_moderno/Descrizione/Capitolo_generale/Lingua_e_scrittura_della_descrizione
- ICCU. 2016b. "Regole italiane di catalogazione. Appendice F – Traslitterazione o trascrizione di scritture diverse dall'alfabeto latino." Last modified September 21, 2016. https://norme.iccu.sbn.it/index.php?title=Reicat/Appendici/Appendice_F
- Inmon, William H. 2005. *Building the data warehouse*. 4th ed. Indianapolis: John Wiley & Sons.
- Ismail, Mohd Ikhwan, and Nurul Azurah Md. Roni. 2010. "Issues and challenges in cataloguing Arabic books in Malaysia academic libraries." *Education for Information* 28 (2-4): 151-163.
- Kim, SungKyung. 2006. "Romanization in Cataloging of Korean Materials." *Cataloging & Classification Quarterly* 43 (2): 53-76. https://doi.org/10.1300/J104v43n02_05
- Kimball, Ralph, Margy Ross, Warren Thorntwaite, Joy Mundy, and Bob Becker. 2008. *The data warehouse lifecycle toolkit*. 2° ed. Indianapolis: John Wiley & Sons.
- Kudo, Yoko. 2010. "A Study of Romanization Practice for Japanese Language Titles in OCLC WorldCat Records." *Cataloging & Classification Quarterly* 48 (4): 279-302. <https://doi.org/10.1080/01639370903338352>
- Levenshtein, Vladimir Iosifovich. 1966. "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* 10 (8): 707-710.
- Li, Yue. 2004. "Consistency versus Inconsistency: Issues in Chinese Cataloging in OCLC." *Cataloging & Classification Quarterly* 38 (2): 17-31. https://doi.org/10.1300/J104v38n02_04
- Molavi, Fereshteh. 2006. "Main Issues in Cataloging Persian Language Materials in North America." *Cataloging & Classification Quarterly* 43 (2): 77-82. https://doi.org/10.1300/J104v43n02_06
- Navarro, Gonzalo. 2001. "A guided tour to approximate string matching." *ACM Computing Surveys* 33 (1): 31-88. <https://doi.org/10.1145/375360.375365>
- Rao, Chaitra, Avantika Mathur, and Nandini C. Singh. 2013. "'Cost in Transliteration': The neurocognitive processing of Romanized writing." *Brain and Language* 124 (3): 205-212. <https://doi.org/10.1016/j.bandl.2012.12.004>

Two Projects and a Thesaurus. Recent Experiences in the Management, Description and Indexing of Oral Sources

Sabina Magrini^(a)

a) Ministero della Cultura

Contact: Sabina Magrini, sabina.magrini@beniculturali.it

Received: 13 April 2021; **Accepted:** 15 July 2021; **First Published:** 15 January 2022

ABSTRACT

The Istituto Centrale per i Beni Sonori e Audiovisivi (ICBSA) has just finished, together with the Università degli Studi di Siena and Università degli Studi di Siena per stranieri, to work on the project “Ti racconto in italiano” which focuses on providing different access points to audio resources collected between 1980’s and 2000’s by ICBSA itself, as part of its mission to document Italian audio and audiovisual culture.

The main aim of the project is to create tools which will enable scholars of social history, art and literature to use these sources as well as providing original material for foreign students to exercise their knowledge of Italian. In order to facilitate access it has been necessary to create finding aids such as indexes and thesauri. For this purpose ICBSA has started a collaboration with the Biblioteca Nazionale Centrale di Firenze and the latter’s Nuovo Soggettario.

This is not the first case of a project by institutes of the Italian Ministry of Culture comprising the use of the Nuovo Soggettario for the indexing of archival materials. Indeed, the Soprintendenza Archivistica e Bibliografica della Toscana has already worked in this direction a few years ago when treating the so-called Straw archives.

KEYWORDS

Nuovo Soggettario; Oral sources; Archives; Indexing.

The aim of this paper is to address a number of significant issues concerning the main theme of this Conference on bibliographic control in the digital ecosystem and mainly:

1. The complex interactions that are becoming common between different areas of knowledge and knowledge management in the bibliographic universe;
2. New ways of indexing documents;
3. The role of Thesauri in digital systems.

To do so, it shall be necessary to concentrate at first on the project “Archivi di paglia” which the Soprintendenza archivistica e bibliografica della Toscana developed around 2014-2016 in collaboration with the Biblioteca Nazionale Centrale of Florence. For those who may not be familiar with the intricacies of Italian cultural administration, the Soprintendenza is a Supervision Agency, the local office of the Italian Ministry for Culture engaged in the protection and valorisation of notified archives and libraries belonging to private individuals or archives and libraries belonging to public (non-State) entities in Tuscany. Object of this project was the census of the companies (and their archives) which produced straw hats in Tuscany in the past or still have connections to that world somehow.

Following, the project “Ti racconto in italiano” shall be illustrated. This is the result of the collaboration between the Istituto Centrale per i Beni Sonori e Audiovisivi (ICBSA, another office of the Ministry which concentrates its activity on the preservation and valorisation of audio and audiovisual heritage), l’Università di Siena (UNISI), l’Università di Siena per Stranieri (UNISTRASI) as well as the Biblioteca Nazionale Centrale of Florence (BNCF). This project is particularly interesting in this context as it bears a great focus on indexing issues and, as it dates back to 2020, it adopts state of the art digital solutions.

These two projects have been chosen as interesting sample cases as they are both recent and quite unique in their sort. Cross-referencing between the worlds of library and archive databases is still relatively uncommon.

As concerns the “Archivi di paglia” project, it is necessary to illustrate at first the context in which the idea of such a research developed.

Archival records have a permanent significance for history, science and culture, as well as for the legal protection of individuals and legal entities. As such they can truly be considered to be a cultural asset.

Amongst the archival records that the Soprintendenza archivistica e bibliografica della Toscana safeguards there are, since the 1970’s, business or company archives. Such archives bear witness to the history, capacities and vision of many big and small intrapreneurs in the local manufacturing industry. Tuscany has been famous since the 18th century for the production and processing of straw. The latter was used to realize the famous ‘cappello di paglia di Firenze’ (viz. the Florentine straw hat), giving work to hundreds of women employed in weaving straw into braids and hats. The tradition has continued somehow until today and some of the older firms are still active: in all, there are around 14 firms, many of which around a century old, which are still producing braids, hats and hat moulds. Their work and tradition has inspired the Museum of Straw in Signa¹, one of the main centres of the production of straw hats. It was in Signa that, for the first time, in 1714 Domenico Michelacci had the idea of starting a new kind of straw crop in order to obtain a thread that was particularly suited for weaving.

¹ <https://www.museopaglia.it/> Accessed June 2021

To celebrate the third centenary of the revolution in straw production and processing in the region after Michelacci's pioneering experiences, the Soprintendenza archivistica opted in 2014 to realize the census of the companies that were or are still active in the sector, to collect data on their archives and to organize and record a series of interviews with business owners, workers and furnishers active in this line (as well as their relatives) in order to have first-hand information on their way of life and work. The interviews (some on video as well) were intended for the Museum at Signa. One of the main aims of the project was also the publication online of the so-called SIUSA descriptions of the archival records of the firms involved. SIUSA is an acronym for the Sistema Informativo Unificato per le Soprintendenze Archivistiche (Unified Information System for the Supervision Agencies). It intends to be the primary access node to non State archival documents, both public and private, which are not kept by State Archives.

The system describes the *archival fonds* according to a multi-level description; the *creators (bodies, people and families)* who produced the documents performing their activities; the persons or bodies who preserve (custody) the archives. General historical, administrative and archival information is provided as well, in order to allow a better comprehension of the context².

The aim of SIUSA is to assure the preservation and the knowledge of these sources and to provide access to them.

It was immediately clear from the perusal of the documents and the examination of the content of the interviews that it would have been essential to dispose of a controlled vocabulary focused on the world of straw processing. In such a way, it would have been possible to choose from a selection of terms in order to index content or to retrieve content through browsing or searching: thus, the SIUSA descriptions and any other works on the archives and interviews would have gained so much in sense, purpose and usability!

The potential of language as a meeting point between libraries, archives and museums was on the other hand a theme of reflection in those years for the Soprintendenza. So much so that in 2012 it had already created the basis of a collaboration with the Biblioteca Nazionale Centrale di Firenze through series of explorative letters and reciprocal declarations of intent.

For this reason, the Soprintendenza archivistica sought the collaboration of the Biblioteca Nazionale Centrale of Florence and specifically of the team behind the Nuovo soggettario³. The Nuovo soggettario viz. the New Subject Index, is the Italian subject indexing tool created by the National Central Library of Florence for the entire system of Italian libraries and, in particular, for those operating in the National Library Service.

Far from being a tool in use only in the Library world, the Nuovo soggettario was and still is open to contributions from other areas of knowledge management and is interoperable with databases of archives and museums as well as available in all standard formats and protocols.

Significantly the potential of this interaction between different worlds was explored on occasion of the conference organized in 2015 by ANAI, the Association of Italian Archivists⁴, MAB *Musei Archivi Biblioteche. Professionisti del Patrimonio Culturale Toscana* as well as the Tuscan Region and hosted by the Soprintendenza itself. On that occasion Emilio Capannelli, one of the archivists

² <https://siusa.archivi.beniculturali.it/cgi-bin/siusa/pagina.pl?RicLin=en> Accessed June 2021

³ <https://thes.bncf.firenze.sbn.it/> Accessed June 2021

⁴ <http://www.anai.org/anai-cms/>; <http://www.mab-italia.org/> Accessed June 2021

of the Soprintendenza, described the first experiences of collaboration between local archivists and librarians⁵.

So, thanks to this collaboration with the National Library, Alessia Artini and Silvia Melloni, the two free lance archivists who worked on the straw archives project under the guidance of the Soprintendenza (and of the archivists Renato Delfiol and Luca Faldi in particular) produced a series of controlled vocabulary terms which have been accepted and adopted by the general Thesaurus that is the main component of Nuovo soggettario system⁶. These terms hence recur in the scientific production concerning the straw archives.

The idea was that of continuing the collaboration between archivists and librarians in order to create other carefully selected lists of words and phrases in order to tag units of information in other 'domains', such as the archival records of other manufacturing sectors. So far, though, this has not occurred, at least as concerns the Soprintendenza archivistica in Tuscany.

A more complex and recent project that develops significantly some of the themes touched by "Archivi di paglia" is that of "Ti racconto in italiano" (fig. 1). This one year long project which has been financed by the General Directorate of Libraries in the Ministry aims at promoting the collections of ICBSA as well as the knowledge of the Italian language and culture abroad. What is it all about? And, first of all, what is ICBSA?



Fig. 1.

⁵ Capannelli 2016, 17-20

⁶ Artini, Benelli and Melloni 2017, 9

ICBSA – as stated before The Central Institute for Sound and Audiovisual Heritage⁷ – was founded in 1928 and was once known as the State Discotheque. Its first collections consisted of a record collection entitled “The word of the Great”, voices collected by Rodolfo De Angelis in the first half of the 1920s. Over the years, documents of folklore, music, history, theater, dance, cinema have been added to this initial nucleus, which represented the first Italian public sound heritage, recorded on the most different media, from the wax cylinders invented by Edison, to records, tapes, videos up to current digital media. ICBSA materials are public and available for consultation via OPAC (Online Public Access Catalogue) with the possibility of listening to the *incipit* of the digitized sound documents and consultation of the accompanying description.

At the beginning of 2020, a collaboration between ICBSA, and the University of Siena (UNISI) and the University for Foreigners of Siena (UNISTRASI) started in order to make a section of the “Historical Voices” collection available to the public.

The materials made available by ICBSA for the realization of the project are 35 interviews on audio files lasting an average of one hour each, carried out between 1983 and 2006. The interviews, chosen by Piero Cavallari (an ICBSA technician) involve prominent personalities from the world of Italian business, art and culture for a total of about 40 hours of recording.

In detail, the corpus of interviews comprises:

- 13 interviews with writers, intellectuals, actors, directors (1983-1989): Elio Filippo Accrocca, poet; Giorgio Bassani, writer; Attilio Bertolucci, poet; Giorgio Caproni, poet; Riccardo Cucciolla, actor; Margherita Guidacci, poet; Luciano Lucignani, film director; Luciano Luisi, poet, writer and reporter; Mario Luzi, poet; Ettore Paratore, Latin scholar; Guglielmo Petroni, writer; Luisa Spaziani, poet; Franca Valeri, actress. (fig. 2)
- 11 interviews with artists (1987-1988): Carlo Belli, intellectual interested in art, architecture, music, archeology, politics; Maria Lai, designer; Carlo Lorenzetti, sculptor in metal; Teodosio Magnoni, painter and sculptor; Elisa Montessori, painter; Alberto Sartoris, architect; Ruggero Savinio, painter and author; Toti Scialoja, painter and poet; Guido Strazza, experimental artist; Giuseppe Uncini, painter and sculptor in iron and cement; Renzo Vespignani, painter, illustrator, set designer and engraver.
- 11 interviews with entrepreneurs (2006): Pia Berlucchi, wine producer; Diana Bracco, health and diagnosis; Filippo Cerruti, tourism, maritime transport, events; Wanda Ferragamo, fashion; Vittorio Ghisolfi, plastic producer; Giorgetto Giugiaro, design; Sergio Giunti, book editor; Ernesto Illy, coffee producer; Steno Marcegaglia, steel producer; Loris Meliconi, house goods producer; Ottavio Missoni, sportsman and fashion designer.

⁷ <http://www.icbsa.it/> Accessed June 2021

The screenshot displays a web interface for a collection titled "Scrittori, Poeti e Attori (STAGE)". At the top, there are four tabs: "Descrizione", "Documenti" (which is selected), "Progetto", and "Diritti". Below the tabs, there is a section titled "Elementi della collezione" with a dropdown arrow. Underneath, it says "Elementi visualizzati 1 - 10 di 13". The main content area lists five audio recordings, each with a small portrait of the author and the following details:

- Accrocca, Elio Filippo - 1987 - Roma**
Data: 1987-02-07
Tipo: Audio
Durata: 01:03:27 (durata temporale)
- Bassani, Giorgio - 1984 - Roma**
Data: 1984-03-28
Tipo: Audio
Durata: 01:10:44 (durata temporale)
- Bertolucci, Attilio - 1984 - Roma**
Data: 1984-03-07
Tipo: Audio
Durata: 01:15:00 (durata temporale)
- Caproni, Giorgio - 1983 - Roma**
Data: 1983-06-06
Tipo: Audio
Durata: 01:09:38 (durata temporale)
- Cucciolla, Riccardo - 1987 - Roma**
Data: 1987-11-28

The bottom of the screenshot shows a Windows taskbar with various application icons and the system clock.

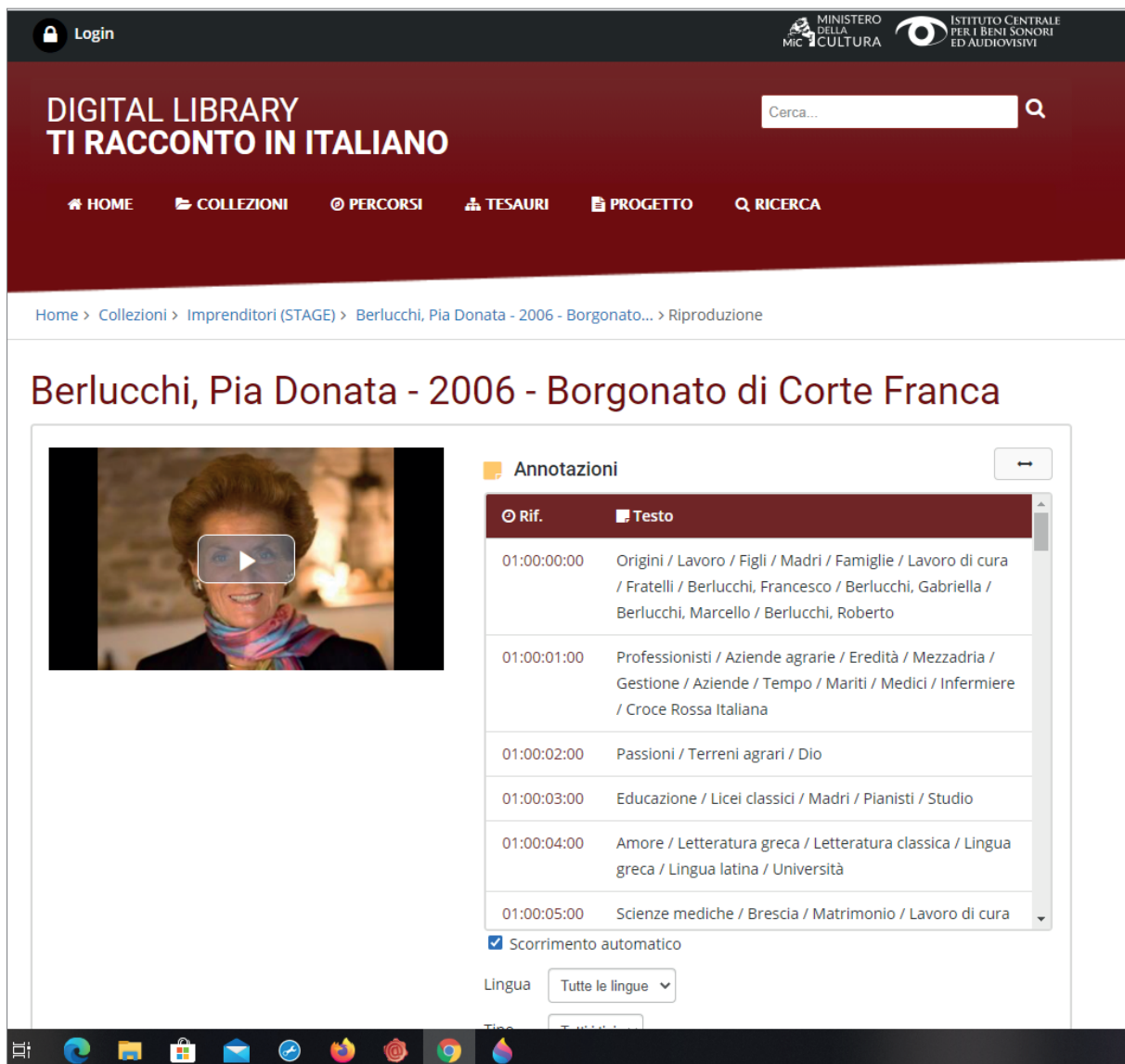
Fig. 2.

As concerns UNISTRASI, ICBSA decided to activate a research grant in order to carry out the didactic adaptation of these sound materials to make them available online to teachers and students of Italian, Level B1-B2. The purposes were two: to bring foreign learners closer to intensive listening to encourage the learning of Italian, but also to open them a world of ideas, stories and emotions linked to some extremely interesting Italian personalities of the second half of the 20th century and through them to offer a cross-section of the Italian culture and society of this period. The work was accomplished by Elena Grifoni under the supervision of Pierangela Diadori (Full Professor of Didactics of Italian for Foreigners).

As Pierangela Diadori has noticed, it is difficult to think that a person with a mother tongue other than Italian and unfamiliar with the Italian culture of the second half of the twentieth century

could listen to interviews lasting over an hour, or read the automatic transcription offered (honestly incomprehensible in many points): not even an Italian native speaker would do it.

The goal was therefore to create a battery of metadata, which could be freely surfed online and associated with the audio files and written files relating to each interview. It all had to be in the form of comprehension or completion exercises with closed answers, to be carried out in relation to the 'listening or reading' of the texts provided, together with keys for self-learning (figs. 3-4).



The screenshot shows the website interface for the Digital Library 'TI RACCONTO IN ITALIANO'. The header includes a login button, logos for the Ministero della Micultura and Istituto Centrale per i Beni Sonori ed Audiovisivi, and a search bar. The main navigation menu contains links for Home, Collezioni, Percorsi, Tesauri, Progetto, and Ricerca. The breadcrumb trail indicates the current page is 'Berluschi, Pia Donata - 2006 - Borgonato... > Riproduzione'. The main content area features a video player on the left and a table of annotations on the right. The table lists time stamps and corresponding topics.

Rif.	Testo
01:00:00:00	Origini / Lavoro / Figli / Madri / Famiglie / Lavoro di cura / Fratelli / Berluschi, Francesco / Berluschi, Gabriella / Berluschi, Marcello / Berluschi, Roberto
01:00:01:00	Professionisti / Aziende agrarie / Eredità / Mezzadria / Gestione / Aziende / Tempo / Mariti / Medici / Infermiere / Croce Rossa Italiana
01:00:02:00	Passioni / Terreni agrari / Dio
01:00:03:00	Educazione / Licei classici / Madri / Pianisti / Studio
01:00:04:00	Amore / Letteratura greca / Letteratura classica / Lingua greca / Lingua latina / Università
01:00:05:00	Scienze mediche / Brescia / Matrimonio / Lavoro di cura

Fig. 3.

The screenshot shows a digital learning interface. At the top left, the title "Pia Donata Berlucci" is displayed in a large, dark red font. Below it, the subtitle "Biografia - comprensione scritta B1" is shown in a smaller, dark grey font. The main content area is divided into two columns. The left column contains a "Domanda 1" section with a date "17 luglio 2018" and a paragraph of text about Pia Donata Berlucci. Below the text is a "Domanda a scelta singola" section with five radio button options. The right column features a navigation menu with icons and text labels: "Introduzione", "Biografia - comprensione scritta B1" (with a sub-menu for "Domanda 1", "Domanda 2", and "Domanda 3"), "Comprensione orale B1" (with a list of topics like "La grande distribuzione", "I successi dei figli", etc.), and "Comprensione scritta B1" (with "Gli studi"). At the bottom of the interface, a Windows taskbar is visible with various application icons.

Fig. 4.

The format had to be such as to allow the insertion of these sets of exercises on the platform, in order to make access faster and offer immediate feedback with keys to the answers. Also a glossary has been realized.

If one thinks of the shortage of didactic audio materials, having this instrument available anywhere in the world online, with the advantage of being free of charge and the possibility of accessing authentic cultural contents, it is easy to understand the importance of this operation⁸.

UNISI's contribution to the project consisted, instead, in the indexing of the text in order to make it interesting for a wider number of potential users and perusable for research purposes. In this

⁸ Diadori 2021

case as well a research grant was activated: the work was accomplished by Cecilia Valentini under the supervision of Silvia Calamai (Associate Professor of Glottology and General Linguistics).

As Valentini and Calamai have pointed out on occasion of the presentation of the project at the CLARIN Annual Conference 2020⁹, the main objective of “Ti racconto in italiano” was to create tools facilitating users search through the collection. Therefore, finding aids such as indexes and thesauri have been realized.

Indexing has been done classifying each segment of the documents at regular time intervals with a label (fig. 3).

The terms used as labels consist of key words and controlled vocabulary and are structured in a specifically created thesaurus, viz. a specialized vocabulary of hierarchically listed words and phrases that indicates a preferred term among synonyms and shows relationships between terms. The use of a thesaurus facilitates retrieval of information and ensures greater consistency in the indexing of documents.

The thesaurus used is naturally that of the Nuovo soggettario¹⁰. The National Central Library of Florence is in fact a partner of the project. Therefore, Anna Lucarelli and her team have followed all phases concerning the choice of the terms and have authorized the use of new terms proposed by Cecilia Valentini and which have now entered the Nuovo soggettario.

Indexing has been done, mainly by Stella Montanari via AVIndexer, a software developed by Davide Merlitti (Informatica Umanistica, Pisa)¹¹ that makes use of SKOS, the Simple Knowledge Organization System recommended by W3C. Also the DublinCore Metadata set has been exploited (figs. 5-6).

Once indexed, the records as well as the thesauri will be published on the internet portal “Ti racconto in italiano” managed by ICBSA. The process is in progress at the moment. This platform is shaped on the digital library *Ti racconto la storia* online since September 2018¹². The latter was conceived by the General Directorate for Archives and the Central Institute for Archives in order to promote the knowledge and use of collections of oral testimonies, stories of life and other audio and audiovisual documentation produced on both analogue and digital media and stored in public institutions, research centers and private associations.

Once online, the perusal of the platform¹³ will clearly show how the resources of the Semantic Web have been exploited in order to create a framework for creating, managing, publishing and searching semantically rich metadata for web resources.

⁹ <https://www.clarin.eu/content/programme-clarin-annual-conference-2020> Accessed June 2021

¹⁰ <https://thes.bncf.firenze.sbn.it/> Accessed June 2021

¹¹ <http://www.informaticaumanistica.com/open-source/avindexer> Accessed June 2021

¹² <https://www.tiraccontolastoria.san.beniculturali.it/> Accessed June 2021

¹³ The foreseen address is: <https://tiracconto.icbsa.it/>

Bassani, Giorgio - 1984 - Roma

▼ Descrizione	▼ Metadati	▼ Trascrizione
Metadati Dublin Core		
DC.Titolo	Bassani, Giorgio - 1984 - Roma	
DC.Soggetto	letteratura / Italia / 1980-1989 / poesia / interviste	
DC.Descrizione	Dopo la nota biografica e bibliografica a cura di Eugenia Tantucci, Bassani passa lettura di versi tratti da varie raccolte, di cui descrive brevemente le forme e i riferimenti. Durante la lettura ribadisce la centralità dell'antifascismo, dell'ebraismo, delle memorie d'infanzia nella sua poetica. Insiste sulla sostanziale uguaglianza fra le sue opere in versi e in prosa e conclude leggendo il capitolo 9 de Il giardino dei Finzi Contini.	
DC.Contributore	Giorgio Bassani (intervistato) / Eugenia Tantucci (intervistatore)	
DC.Data	1984-03-28	
DC.Tipo	Audio	
DC.Formato	01:10:44	
DC.Identificatore	ICBSA:DDS1588511	
DC.Fonte	1 Nastro (bobina aperta) (120 min. ca.); 7 1/2 in. per sec. (19 cm.), Elettrica/analogica, Stereofonico, Originale, AGFA PER 525	
DC.Relazione	http://polodds.dds.it/opac2/DDS/dettaglio/documento/DDS1588511	
Metadati MPEG-7		
Titolo	Bassani, Giorgio - 1984 - Roma	
Riassunto	Dopo la nota biografica e bibliografica a cura di Eugenia Tantucci, Bassani passa lettura di versi tratti da varie raccolte, di cui descrive brevemente le forme e i riferimenti. Durante la lettura ribadisce la centralità dell'antifascismo, dell'ebraismo, delle memorie d'infanzia nella sua poetica. Insiste sulla sostanziale uguaglianza fra le sue opere in versi e in prosa e conclude leggendo il capitolo 9 de Il giardino dei Finzi Contini.	

Fig. 5.

Metadati MPEG-7	
Titolo	Bassani, Giorgio - 1984 - Roma
Riassunto	Dopo la nota biografica e bibliografica a cura di Eugenia Tantucci, Bassani passa lettura di versi tratti da varie raccolte, di cui descrive brevemente le forme e i riferimenti. Durante la lettura ribadisce la centralità dell'antifascismo, dell'ebraismo, delle memorie d'infanzia nella sua poetica. Insiste sulla sostanziale uguaglianza fra le sue opere in versi e in prosa e conclude leggendo il capitolo 9 de Il giardino dei Finzi Contini.
Produttore	Discoteca di Stato
Intervistatore	Tantucci Eugenia
Intervistato	Bassani Giorgio
Luogo	Roma
Data	1984-03-28
Genere	Intervista
Forma	Serie
Soggetto	letteratura / Italia / 1980-1989 / poesia / interviste
Lingua	it-IT
Durata	01:10:44
Parole chiave (SKOS)	Bologna / Università degli Studi di Bologna / Longhi, Roberto / Tantucci, Eugenia / Famiglie / Ferrara / Città / Ispirazione poetica / Attaccamento / Letteratura / Politica / Attività clandestina / Giovani / Intellettuali / Partito d'azione / Arresto / Antifascismo / Armistizio dell'8 settembre <1943> / Roma / Pubblicazione / Poesia / Bassani, Giorgio, Storie dei poveri amanti e altri versi / Bassani, Giorgio, Te lucis ante / Bassani, Giorgio, Un'altra libertà / Bassani, Giorgio, L'alba ai vetri / Redattori / Periodici / Botteghe oscure <Periodico> / Paragone <Periodico> / Bassani, Giorgio, Cinque storie ferraresi / Bassani, Giorgio, Gli occhiali d'oro / Bassani, Giorgio, Il giardino dei Finzi-Contini / Premio Viareggio / Bassani, Giorgio, Dietro la porta / Bassani, Giorgio, L'airone / Premio Campiello / Bassani, Giorgio, Le parole preparate e altri scritti di letteratura / Bassani, Giorgio, L'odore del fieno / Racconti / Saggi / Bassani, Giorgio, Epitaffio / Bassani, Giorgio, In gran segreto / Bassani, Giorgio, Il romanzo di Ferrara / Dediche / Bassani, Giorgio, In rima e senza / Premio Bagutta / Dortmund / Premio Nelly Sachs / Traduzioni / Narrativa / Sintassi / Lessico / Risentimento / Dolore / Vita / Storia / Solitudine / Discriminazione razziale / Religiosità / Laicismo / Emarginazione / Società / Contestazione / Dubbio / Antenati / Libertà / Bassani, Giorgio / Rima / Poeti / Versi / Alessandrini / Dodecasillabi / Settenari / Strofe / Iscrizioni / Epitaffi / Lapidi / Cimiteri / Case / Jahier, Piero /

Fig. 6.

References

Artini, Benelli and Melloni 2017. *Archivi di paglia. Gli archivi del distretto industriale della paglia in Toscana*, edited by Alessia Artini, Angelita Benelli, Silvia Melloni, Firenze: Edizioni Polistampa.

Capannelli, Emilio 2016. “Prime esperienze di collaborazione tra archivisti e bibliotecari” In *Il nome delle cose. Il linguaggio controllato come punto di incontro tra archivi, biblioteche e musei. L’esperienza del Gruppo linguaggi di MAB Toscana*, edited by Francesca Capetta, 17-20. Accessed June 2021. http://www.ilmondodegliarchivi.org/images/Quaderni/MdA_Quaderni_n1.pdf

Diadori, Pierangela 2021. “Ti racconto in italiano: voci del ’900 per imparare l’italiano L2. Progetto ICBSA-UNISTRASI di didattizzazione di interviste sonore a personalità del ’900.” In *La Nuova DITALS risponde 3*, edited by Pierangela Diadori, Caterina Gennai, Elena Monami, in press. Roma: Edilingua.

The bibliographic control of music in the digital ecosystem. The case of the Bayerische Staatsbibliothek (BSB)

Klaus Kempf^(a)

a) Independent consultant – formerly Bayerische Staatsbibliothek

Contact: Klaus Kempf, klauskempf@gmx.de

Received: 19 May 2021; **Accepted:** 12 June 2021; **First Published:** 15 January 2022

ABSTRACT

The BSB's music department (entrusted since 1949 with the management of the national information service on music) is one of the largest music libraries in the world in terms of the size and quality of its collection, but also in terms of the breadth and depth of its collection acquisition policy. The various materials are widely catalogued and indexed in a very articulate way, using a wide range of catalogues and according to specific rules. The BSB currently uses the RDA and MARC21, according to national policies.

The Gemeinsame Normdatei (GND), the authority files of the German-speaking library world, are used both in cataloguing and in subject classification. The GND is nowadays used even outside the library world by archives, museums and other kinds of institutions, as well as for the cataloguing of websites.

The BSB participates in the RISM (Répertoire International des Sources Musicales) international online catalogue of music sources, and, together with the Staatsbibliothek zu Berlin, manages its OPAC.

The presentation will describe these projects, as well as the cataloguing workflow, the application of the RDA in specific cases, the special rules (and cataloguing system) for personal archives and musical legacies (RNA), and finally the futuristic service 'musiconn'. This last service is included in the national service for music information Fachinformationsdienst Musikwissenschaft and has been developed by the BSB: it offers the possibility to search by melody, as part of a project based on Optical Music Recognition (OMR), a software tool that allows automatic recognition of compositions after they are printed.

KEYWORDS

Cataloging music sources in Germany; German authority files in musicology and music sources; Digitization of music sources.

The Bayerische Staatsbibliothek as a big music research library

Subsection title

The Bayerische Staatsbibliothek (BSB) is not only the central state library (national library) of the Free State of Bavaria, but also or in particular a big research library. It disposes of world wide well known and recognised special collections in a couple of science disciplines. One of them regards music and musicology. The library hosts in its stacks 455.000 music editions, 72.000 music manuscripts, 330 composer archives (personal papers), 93.000 non book material/sound carriers, in particular discs and CDs, 164 000 books and journals about music and musicology. Since 1949 the BSB is part of the national special collection programme, especially music, cofinanced by the German Research Society (DFG) and since 2014 the library is responsible for music and musicology within the framework of the Specialised information services programme (FID) also cofunded by the DFG.

Cataloging & metadata management in music and musicology collections

Following the principle that cataloging aims traditionally on two major objectives: on the one hand side on the specificity of the object/material regarded: on the other hand side on the needs (and desires) of the (potential) user leads in the field of music and musicology – at least in the German speaking world – to an especially varied and contemporarily particularly profiled cataloging/metadating in a relatively wide range of different catalogs.

Since the introduction of the online catalogs (OPAC) the German libraries use more and more the in the meanwhile well established standards,

- formed by the national/international cataloging rules together with the special guidelines for music (in former times RAK /Musik now a days RDA and the music specific guidelines – <https://wiki.dnb.de/display/RDAINFO/Arbeitshilfen>); and
- standardised data formats (MAB2 and MARC 21; in the case of personal papers also EAD);
- today libraries use for cataloging, and subject heading (the same) authority files. In Germany in both cases they use the Integrated Authority File (GND).

The integrated authority file (GND)

The Integrated Authority File (Gemeinsame Normdatei – GND: https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html;jsessionid=935B36EDCD89249E62A%201BA3000574759.internet531) is a service facilitating the collaborative use and administration of authority data. These authority data represent and describe entities, i.e., persons, corporate bodies, conferences and events, geographic entities, topics and works relating to cultural and academic collections. Libraries in particular use the GND to catalog publications. However, archives, museums, cultural and academic institutions, and researchers involved in research projects are also increasingly working with the GND. Authority data make cataloging easier, offer definitive search entries and forge links between different information resources. Every entity in the GND features a unique and stable identifier (GND ID). This makes it possible to link the authority data

with both each other and external data sets and web resources. This results in a cross-organizational, machine-readable data network.

Cataloging of music editions, books, and audio/sound carrier in Germany

Cataloging of music editions, books and audio/sound carriers in Germany is normally done on a regional level via the academy library system dominating regional and intraregional library networks, however periodicals are cataloged in a nationwide, even transnational database. In the case of the BSB music editions, books (monographs) and audio/sound carriers are cataloged in the Union Catalog (Verbundkatalog – B3kat) of the Bavarian Library Network (BVB): <<https://www.bib-bvb.de/>>. But the periodicals, journals, year books and so on, are cataloged – like the other libraries are doing – in the German National Periodical Catalog (ZDB): <<https://www.zeitschriftendatenbank.de/startseite>> on which are participating also the library systems of Austria and the German Switzerland. By adding a shelfmark at the cataloging record in the union catalog and the national periodical catalog the concerned record is replicated/duplicated – in real time – in the regarded local OPAC: <<https://opacplus.bsb-muenchen.de/metaopac/start.do>>.

In a different way are handled the music sources. They are primarily cataloged worldwide in cooperation via RISM <<https://rism.info/index.html>>. The Répertoire International des Sources Musicales (RISM), International Inventory of Musical Sources, is an international, non-profit organization that aims to comprehensively document extant musical sources worldwide: manuscripts, old music editions, writings on music theory, and libretti that are found in libraries, archives, churches, schools, and private collections. The RISM Catalog of Musical Sources contains over 1.2 million records and can be searched at no cost. RISM was founded in Paris in 1952 and is the largest and only global organization that documents written musical sources. RISM records what exists and where it can be found.

The cataloging happens decentrally via an international cooperation following specific cataloging rules which are primarily oriented on musicological criteria. This catalog in the case of handwritten material offers access to its content also via the cataloging of the so called music incipits. Moreover, this catalog is using an own rather detailed (meta)data format. At least the German editorial staff is also using systematically the (German) Authority File, the GND. The central cataloging tool, the data base, is hosted by the Staatsbibliothek zu Berlin (SBB). The (RISM) OPAC is managed and maintained by the BSB. The BSB-OPAC is updated every half a year with the new records cataloged in the RISM-Catalog via the Bavarian Network Union Catalog (B3Kat).

Cataloging of personal papers (composer archives), publisher archives and manuscripts

Kalliope is a Union Catalog for collections of personal papers, manuscripts, and publishers' archives and the National Information System for these material types (<<https://kalliope-verbund.info/de/index.html>>). Founded by the Berlin State Library – Prussian Cultural Heritage with financial support from the German Research Foundation (DFG) in 2001, Kalliope superseded the Central Register of Autographs (Zentralkartei der Autographen, ZKA), which was established in 1966.

The joint cataloging in Kalliope is based on established archival and librarian description and cataloging guidelines and relies heavily on authority control processes. Kalliope is therefore not just another data aggregation service, but rather a digital environment that establishes and provides new instruments and processes to create, modify, and to access data about personal papers dispersed in many libraries, archives, and museums. But using Kalliope means parallel cataloging and parallel offer of data access. Until today there is no connection, neither an interface nor a data transfer (replication) between the Kalliope data base and the union catalogs of the various regional library networks or the single local OPACs.

Kalliope: History, Development, State of the Art

The initial data base of Kalliope was formed by 1,2 million catalog cards of the ZKA that had been provided by 450 institutions over a period of more than 30 years. The conversion of these cards into a machine-readable format was completed in 2006. Moreover, the catalog service was extended step by step to provide access to collections of personal papers in Austria and Switzerland as well as personal papers of persons from German speaking countries kept in libraries and archives abroad, particularly in the United States of America. Since 2001 cultural heritage organizations can make full use of a client-server based cataloging application including full access to the Integrated Authority File (Gemeinsame Normdatei, GND) of the German National Library. The Union Catalog takes an active role in the operation of this national cataloging resource and adds greatly to it by identifying entities that are only known via unique materials as are described in Kalliope.

The cataloging client conforms to the German Guidelines for the description of personal paper and manuscript collections (RNA – Regeln für die Erschließung von Nachlässen und Autographen) which are in turn compatible with ISAD(G) – General International Standard Archival Description. As of May 2015, 102 organizations use the Kalliope cataloging client – compared to 54 in 2010. Additionally, standardized data (EAD – Encoded Archival Description) from local applications can be made available for retrieval in Kalliope. Currently the database provides access to 19,300 collections with a total of more than 3 million units of description originating from more than 950 institutions, including letters, manuscripts, personal documents, albums, diaries, lecture notes, photographs, posters, movies, screenplays, music editions and even some famous ringlets. The database includes around 600,000 name records, 253,000 of which describe individualized persons distinguished by a unique identifier of the GND, and more than 90,000 records of corporate bodies, with 24,000 of these having a unique identifier of the GND.

Kalliope and bibliographic control

The Kalliope Union Catalog is committed to comply with standards (guidelines, file formats, authority files, ISO norms) of the library and archival community: Encoded Archival Description (EAD): XML schema for encoding archival finding aids; Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC): XML schema for encoding (archival) authority records; GND – Integrated Authority File of the German National Library: uniquely referenced vocab-

ulary for entities such as persons, corporate bodies, places, and subject headings Guidelines for the Description of Personal Paper and Manuscript Collections (Regeln für die Erschließung von Nachlässen und Autographen, RNA); ISO 15511: International Standard Library Identifier and Related Organisations (ISIL): unique identifier code for scientific and cultural heritage organizations; ISO 3166: Codes for Names of Countries, dependent territories, special areas of geographical interest and their principal subdivisions: used for assigning persons and corporate bodies of the integrated authority file to main geographical area; ISO 629-2: Codes for the representation of names of languages – Part 2: Alpha-3 code: used to describe the language within a unit of description. The Guidelines for the Description of Personal Paper and Manuscript Collections are well established and applied by libraries, archives, museums, and similar organizations in Austria, Germany, and Switzerland, and are compatible with the principles of archival description outlined in the General International Standard Archival.

The new Special information service (FID) for music/musicology – musiconn

The introduction of information portals or platforms is a web conform way to put order and structure in rather heterogenous data, but concerning the same scientific discipline; in particular they offer normally a discipline specific unique search possibility on heterogenous information sources. In German this way of presenting information is often called Sekundär-erschließung. The service musiconn.search (<https://www.musiconn.de/>) offers access to 19 relevant data bases/catalogues and other, also fulltext online sources with 6,5 millions single items. A special service of musiconn is the so called melodies search, the musiconn.scoresearch. The prototypical application was developed by the BSB itself, promoted and financially supported by the DFG. The software tool is based on the principles of Optical Music Recognition (OMR) and allows the automatical recognition of melodies in selected digitized music sheets. Actually the melody search is possible in the compositions of the following composers (<https://scoresearch.musiconn.de/ScoreSearch/about>).

Conclusion

The start of the online cataloging pushed the standardization in general and involved a consequent usage of authority file controlled terms in cataloging as well as in subject heading. This applies in principle also to information material regarding music and musicology. In the field of music and musicology cataloging traditionally there is a strong input from the research community itself. The organizational platform for that is RISM, an international body which has established just in the early 50ies its own cataloging database. Today the database is a publicly accessible catalog for the registering (and cataloging) of music sources, in particular manuscripts and old music editions as well as libretti. The RISM-Catalog has its own rules and is based on its own data format. This catalog offers access to its content also via the cataloging of the so called music incipits.

In the German speaking world existing online platform for the cataloging of personal papers & autographs and giving access to them in the internet, called Kalliope, is also used for the catalog-

ing of relevant material in music and musicology. Last but not least the recently introduced special information service (FID) for music, musiconn, offers with musiconn.scoresearch the possibility to find melodies in selected and via the portal accessible digitized music editions. Score search is still a work in progress and not error-free, but it is a decisive step towards a machine learning approach in the field of musicology. It is similar to a full text search in text based sciences and in a near future it can become an very interesting service also for the average user.

Riviste digitali e digitalizzate italiane (RIDI): a reconnaissance for the national newspaper library

Fabio D'Orsogna^(a), Giulio Palanga^(b)

a) Biblioteca nazionale centrale di Roma, <http://orcid.org/0000-0001-9578-8715>

b) Biblioteca nazionale centrale di Roma, <http://orcid.org/0000-0001-9737-2529>

Contact: Fabio D'Orsogna, fabio.dorsogna@beniculturali.it; Giulio Palanga, giulio.palanga@beniculturali.it

Received: 14 April 2021; **Accepted:** 31 May 2021; **First Published:** 15 January 2022

ABSTRACT

The article presents a reflection born from a reconnaissance (named RIDI, Riviste digitali e digitalizzate italiane) launched in December 2019, on online open access journals and digitalization of previously printed publications, which are not always considered as unitary bibliographic elements. It highlights the increasingly urgent need to offer its users not only the physical heritage of the library but also the entire world of open-access digital publications available on the web. Starting from an overview of the state of the art of Italian open access periodicals, both digital natives and continuations or parallel editions of previously printed publications, it offers some examples of bibliographic records already present in the national OPAC of SBN (Italian union catalog), related to publications available with both printed and digital editions. It illustrates the main Italian and international digital libraries, highlighting the problems of coordinating the various initiatives to improve the quantitative and qualitative offer of products. The directory, from which a database integrated with the portal of the Digital Newspaper Library of the National Central Library of Rome will originate, will allow direct access to resources through multiple search fields. The prototype of a super-record of the Work will provide all the elements for the standardization of images, data, metadata, bibliographical histories of publications, to build the national digital newspaper library of the future.

KEYWORDS

Digital libraries; Digital newspaper library; Open access serials; RIDI; National Central Library of Rome.

This paper takes its cue from a reconnaissance called RIDI (Riviste digitali e digitalizzate italiane) launched in December 2019, on those bibliographic realities often not considered in a unified way on our OPACs, such as open access online journals and digitalization of previously printed publications. The need, more than 20 years after the first digitization projects that involved Italian libraries, is to offer its users not only the physical heritage of its library but also the whole vast world of open access digital publications available on the web.

The paper will illustrate the state of the art on Italian open access periodicals, either digital natives, continuations, or parallel editions of previously printed publications, then it will propose some examples of bibliographic records already present in the Italian national catalog, the SBN (Servizio bibliotecario nazionale) OPAC,¹ related to publications with both printed and digital editions, then it will provide a prototype that can provide the elements for a qualitative standardization of images, data, metadata, bibliographic histories of publications, to build the national digital newspaper library of the future.

Better late than never

In the Wiki on Open Access in Italy, a portal that records news and information about the movement at the national and international level, under the heading *Riviste italiane OA* (Italian OA Journals), this communication appears: “At this time the requested page is empty. You can search for this title in other pages of the site or search in related registries, but you do not have permissions to create this page”.²

This first Italian directory, which does not presume to be exhaustive, arrives with some delay and tries to fill a gap. It is now about 20 years that the main faculties and university departments in the world have begun to organize themselves to offer their journals online. In 2000, the Cato Institute, a temple institution of US liberalism, dedicated its annual conference to the question of which of the two paradigms – intellectual property or open access – would dominate the economy of the future.³ In 2003, many scientific institutions signed the Berlin Declaration on Open Access to Scientific Literature. In Italy, in November 2004, the Berlin Declaration was followed by the Messina Declaration, joined by about thirty universities.⁴

Universities began to organize themselves by creating dozens of university presses. In 2009 the UPI Coordination was established, which in 2018 became the *Associazione Coordinamento delle University Press Italiane*.⁵

¹ <https://opac.sbn.it/opacsbn/opac/iccu/free.jsp>

² https://wikimedia.sp.unipi.it/index.php?title=Riviste_italiane_OA

³ Carlo Formenti, *Corriere della sera*, 20 novembre 2000, p. 27; see also http://www.treccani.it/vocabolario/open-access_%28Neologismi%29/

⁴ Bologna, Brescia, Calabria, Firenze, Foggia, Genova, Insubria, Lecce, Messina, Milano, Milano Bicocca, Milano Politecnico, Milano Vita-Salute San Raffaele, Modena, Molise, Napoli Federico II, Napoli L'Orientale, Napoli Partenope, Padova, Palermo, Parma, Piemonte Orientale, Roma LUMSA, Roma Tor Vergata, Roma III, Siena, Torino, Trieste, Trieste SISSA, Tuscia, Venezia IUAV, and Istituto Italiano di Medicina Sociale di Roma.

⁵ The association aims to study and deepen the issues related to the positioning, the function, and the promotion of university publishing and popular science as well as the possibility to participate in national and international calls for funding of publishing projects. 13 university publishing houses publish 25 open access journals.

Reasons for growth

We are still far from the total replacement of the printed page by the web page, but the steady growth in the number of magazines appearing online is no less real.⁶ The reasons for this success are very simple: plenty of space reduced publication costs and, above all, ease of access anywhere with just the availability of a network connection.

Underlying the success of the Open Access Initiative are two instances:

1. increase dissemination, visibility, and impact of scholarly literature through publication in open, online, institutional, and disciplinary repositories;
2. to counteract the rising prices of academic journals with alternative models of scholarly communication.

For many small businesses, bearing the economic burden of printing and shipping magazines has become unsustainable and is often the motivation to publish only in digital format. This transformation, feared by many, which represents a surrender to the affordability of digital, often also allows a qualitative leap and a broadening of the horizons of publications. Online publication can enhance the characteristics of periodicals and allow readers to navigate the texts in a simpler, more agile, and sometimes interactive way.

There are two models for sustaining management costs and remaining adherent to the philosophy of free access; the model centered on financing by consumers of content (demand-side) and that financed by content producers through sponsorship, donations, fundraising (supply-side). The main supply-side model is that of the Article Processing Charge (APC), better known as the author-pays model, which provides for the payment by the authors of articles accepted for publication of a contribution, which can reach in some cases up to \$ 2,500, while for the contributors of articles from poor or developing countries, the publication is free.⁷

Legal deposit of digital resources in Italy

Legal deposit, i.e. the compulsory delivery of publications to depository institutions by the subjects envisaged by Italian Law no. 106 of April 15, 2004, and Presidential Decree no. 252 of May 3, 2006, is the regulatory instrument that allows the collection and preservation of the various publications in national and regional archives. The law also deals with native digital publications (born-digital).

Two significant experiences were born as a result of the law.

CNR SOLAR (Scientific Open-access Literature Archive and Repository) is a database of scientific publications, established in 2006, aimed at creating an archive of Italian products of science and research, using also the Legal deposit of digital publications. In the context of the mission entrusted to the CNR Central Library by the Law 2004/106 and by the Presidential Decree 2006/252, the

⁶ The Directory of Open Access Journals (DOAJ) listed 2,100 academic-level journals in 2006; as of April 3, 2021, there are 16,146.

⁷ A clear and comprehensive account of the costs of the Open Access publication process can be found in Technical report #1 (2018) from CNR Bologna: Mangiaracina, Silvana and Cristina Morroni. 2018. *Quanto costa l'accesso alle pubblicazioni scientifiche nell'era dell'Open Access? : una prima analisi delle pubblicazioni nel CNR*. Bologna: Biblioteca Area della ricerca di Bologna CNR. <https://zenodo.org/record/1247497#.XoC-JKPOPkU>

legal deposit is aimed at constituting the Italian archive of scientific publications and at realizing national bibliographic services of information and access to the documents subject to legal deposit. Legal deposit in SOLAR is realized through:

1. self-archiving by the author(s), who must make sure of the actual conditions of use and dissemination of the version of the deposited work, previously agreed upon with the publisher and/or producing institution;
2. specific agreements between the CNR Central Library and the publisher and/or the producing institution of the publications. In this case, the deposit may be made by the Central Library itself or by the publisher/producing institution.

The resources in SOLAR can be full-text open access or limited access, i.e. the metadata are still accessible, while it is necessary to contact the CNR Central Library for full-text resources.

Magazzini Digitali is the Italian project for digital legal deposit, launched on July 14, 2011, with the signing of an agreement between the Ministry of Cultural Heritage and the Presidents of the most representative associations of the publishing industry: AIE, FIEG, USPI (later joined by MEDIACOOOP and ANES).

The purpose of the agreement was to promote the experimentation of the legal deposit of born-digital works in the National Central Libraries of Rome and Florence and, limited to the backup copy, in the Biblioteca Nazionale Marciana of Venice.

The experimentation lasted 3 years, starting in 2012. After this period, a shared and efficient system of legal deposit should have been outlined and, in particular, the procedures related to digital works should have been defined through the issuing of a specific regulation.

The trial ended on December 31, 2014. Magazzini Digitali continued to receive subsequently few publications covered by the agreement in the 2012-2014 Conventions, receiving at sperimentazione@depositolegale.it requests for voluntary membership, while waiting for a final regulation and trying to cope with requests based on the few resources available.

The budget is insufficient, as is, more generally, the response of Italian cultural institutions to the preservation of this type of publications, which will inevitably be lost if no concrete and adequate action are taken to deposit them in national archives as is the case for printed publications.

Contents and purpose of RIDI

RIDI (Riviste digitali e digitalizzate italiane) is a repertory conceived as a work in progress that already contains the bibliographic records of about 12,000 Italian journals, compiled according to the standards of the SBN cataloguing guide,⁸ with their URIs,⁹ available on the Internet for free access. All journals that require subscription and registration for a fee are excluded.

There are two main reasons for this choice: the first is practical and is based on the consideration that online journals now represent an enormous quantity, probably more than that of printed journals, which makes bibliographic control almost impossible, as Mauro Guerrini predicted in

⁸ https://norme.iccu.sbn.it/index.php/Guida_moderno

⁹ https://it.wikipedia.org/wiki/Uniform_Resource_Identifier

1999.¹⁰ The second is more exquisitely librarian: both from the cataloguing point of view since it gives a standardized account of bibliographical descriptions that would otherwise be absent from the web and, above all, from national and local OPACs of resources that are unknown to catalogs; and from the point of view of the preservation of printed copies of dual-track publications (paper and online), since it would be possible to exclude from ordinary consultation all those resources that are freely available on the Internet and of which information has been given in catalogs.

The repertory is currently complete only for digital journals and is being completed for the part concerning journals digitized from paper format. The final goal will be to create a single repertory containing also the ever-growing world of resources born in print and digitized later, as a result of public and private digitization campaigns in recent decades. The digital recovery of a printed past, among other things, is present in many journals that are entering open access after a long paper season and represent an attempt to progressively provide all the published material in a single digital archive. Think, for example, of what Banca d'Italia has done in the last 10 years (it has 97 open access publications in its portal) which has made an enormous recovery of its historical publications.¹¹

The search for digitized journals began in April 2020. The work will involve the analytic cataloguing of 73 Italian digital libraries surveyed. As for the type of resources, RIDI includes:

- a) Italian native digital journals, which are one-tenth of the total;
- b) journals published in mixed form, in print and online. Of the publications of this second type, the description of the printed part has also been given, highlighting all the connections between the two forms of publication;
- c) digitized Italian journals.

Intending to find titles even outside the academic circuits, we have therefore given an account of the relations, more and more numerous and frequent, between printed publication and open access publication within the history of the same publication. This allowed us to adequately reconstruct the historical evolution of many journals, also from a cataloguing point of view, to offer the OPAC, in the case of Italy the SBN OPAC, the possibility of providing adequate information on their publishing history and to start the cataloguing of the online issues in SBN, both by intervening in the area of notes and URIs¹² and by creating new bibliographic descriptions linked to the descriptions of the printed editions. During the editing of this catalog, numerous bibliographical descriptions of online resources not yet present were created on the SBN OPAC.

Take the case of *Giornale di gerontologia*, a prestigious journal published for sixty years by the Italian Society of Gerontology and Geriatrics. In 2013, it ceased its print publication. A laconic note informs SBN OPAC users that since 2014 it is published only online. Actually, the journal

¹⁰ “La proliferazione incontenibile delle basi di dati ad accesso remoto rende evidente come sia oggi più che mai illusorio il controllo bibliografico universale [...] la biblioteca può pensare di descrivere solo le risorse elettroniche di proprio interesse [...] selezionando le risorse in modo piuttosto stretto”. Guerrini, Mauro, “Catalogare le risorse elettroniche: lo standard ISBD(ER)”, *Biblioteche oggi*, 17 (1999), no. 1, 62.

¹¹ See <https://www.bancaditalia.it/pubblicazioni/relazione-annuale/index.html> for the annual reports of Banca d'Italia governors from 1894 to 2019.

¹² In this regard, see the ICCU note that established the rules for including in the General Content Notes of the ISBD the link to the digitized copy of the copy not owned by the operating library. <http://polonap.bnnonline.it/index.php?it/21/news-ed-venti/46/link-alla-copia-digitalizzata-dellesemplare-non-posseduto-dalla-biblioteca-operante-regole-per-linserimento>

only retrieves online a few previous years and since 2016 it changes its title. Users have no news of this. RIDI offers this fundamental information to reconstruct the entire bibliographic history of the periodical.

***Giornale di gerontologia** : organo ufficiale della Società italiana di gerontologia e geriatria. – Anno 1, n. 1/2 (gen.-feb. 1953)-anno 61, n. 6 (dicembre 2013). – Firenze : L. Macrì, 1953-2013. – 61 volumi : ill. ; 25 cm. ((Mensile; poi bimestrale. – Il formato varia in 30 cm. – La casa editrice varia: Pisa : Pacini. – BNI 1953-5821. – ISSN 0017-0305; poi 0367-4533. – Dal 2014 solo on line. – CFI0353910

Ha come supplemento: *Giornale dell'arteriosclerosi

***Giornale di gerontologia** : organo ufficiale della Società italiana di gerontologia e geriatria. – Anno 58, n. 1/2 (gen.-feb. 2010)-anno 63, n. 4 (dicembre 2015). – Pisa : Pacini, 2014-2015. – 34 File PDF. ((Bimestrale; trimestrale nel 2015. – ISSN 2035-021X. – Disponibile in Internet all'indirizzo: <http://www.jgerontology-geriatrics.com/issue/archive>

Continua con: *JGG : *Journal of gerontology and geriatrics

***JGG : *Journal of gerontology and geriatrics** : official journal of the Italian Society of gerontology and geriatrics. – Vol. 64, 01 (2016)-. – Pisa : Pacini, 2016-. – File PDF. ((Trimestrale. – ISSN 2499-6564. – Disponibile in Internet all'indirizzo: <http://www.jgerontology-geriatrics.com/issue/archive>

Autore: Società italiana di gerontologia e geriatria

Soggetto: Geriatria – Periodici; Gerontologia – Periodici

Classe: D618.97005

Giornale di gerontologia *bibliographic record on RIDI*

RIDI is ordered alphabetically by title. The source of the bibliographical information is CAPUS (Catalogo delle Pubblicazioni in Serie possedute dalla Biblioteca nazionale centrale di Roma).¹³

In order to give full visibility to this repertory and to expand its search possibilities, it will be necessary to create a database, which will allow direct access to the resources through multiple

¹³ CAPUS is a catalog edited by Giulio Palanga and published in 2019, containing the complete collection of all periodicals and newspapers owned by the National Central Library of Rome. Divided into twelve volumes, the first two volumes contain the index of titles and the index of authors and subjects and constitute the guide by which to navigate the catalog knowing a title, an author, or a subject to search for. The other ten volumes contain the bibliographical records of over 72,000 periodicals, divided into 53 sections, which contain a unique alphanumeric code that refers to a single access point in the catalog, with all the history and editorial changes of the publication, without the need to navigate through the various titles that periodical publications often adopt. <http://www.bncrm.beniculturali.it/it/325/archivio-news/3259/>

search fields.¹⁴ All the descriptions, however, already allow a hypertextual link to digital or digitized resources.

This first compilation of the catalog, completed on April 7, 2020, includes 11,706 bibliographic records. We started by retrieving information from CAPUS, where over twelve years of editing, links, and URIs with online publications of printed journals were gradually reported. This first reconnaissance has allowed us to find the journals contained in the 66 main Italian open access publishing platforms that have been analytically catalogued, verifying the correctness of the URI links, leaving out those no longer traceable on the web. The vast majority of these titles are also present in the two most important international sources, the ISSN portal with 1,028 titles,¹⁵ the DOAJ (Directory of Open Access Journal) with 461 titles,¹⁶ and, for Italy, Magazzini digitali with 113 titles.¹⁷

Open access, digitization, and bibliographic control

Online publications are often accompanied by the digital retrieval of issues published in print. Sometimes, this can happen by chance, but it is now possible to reconstruct and document the history of a publication through the various phases of its editorial policy, which almost always start from the printed text and end with the online publication and the digitalization of previous years.¹⁸

See, for example, the digitization of the entire archive of Radiocorriere, a weekly magazine that was the official organ of RAI for seventy years, from 1925 to 1995. With all the schedules and articles of the newspaper, it is possible to reconstruct the bibliographic (and political) history of the publication. An enormous amount of unpublished material, which represents a unique testimony and an exclusive source for contemporary historiography, not only of the media. It is one of the treasures recovered by Teche RAI and made available to the network free of charge.

¹⁴ For example, in addition to the title, one could include search fields for author, subject, DDC, and provide a permalink of the resource and a permalink of the description in the catalog, as well as the holdings of the printed resource and the holdings of the online or digitized resource.

¹⁵ For the alphabetical list of Italian periodicals see: [https://portal.issn.org/?q=api/search&search\[\]=MUST=country=ITA&search_id=7564722&sort=sort.title](https://portal.issn.org/?q=api/search&search[]=MUST=country=ITA&search_id=7564722&sort=sort.title). Not all titles belong to actual periodicals. As is well known, the ISSN is attributed both to periodicals and to series and monographic series.

¹⁶ For the alphabetical list of Italian periodicals see: https://doaj.org/search/journals?source=%7B%22query%22%3A%7B%22filtered%22%3A%7B%22filter%22%3A%7B%22bool%22%3A%7B%22must%22%3A%5B%7B%22terms%22%3A%7B%22index.country.exact%22%3A%5B%22Italy%22%5D%7D%7D%5D%7D%7D%2C%22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%7D%7D%2C%22size%22%3A%50%2C%22sort%22%3A%5B%7B%22created_date%22%3A%7B%22order%22%3A%22desc%22%7D%7D%5D%7D

¹⁷ <http://www.depositolegale.it/journals/>

¹⁸ *Lucifero : periodico democratico-radical. - [S. l. : s. n., 1870]- (Ancona : Tip. sociale). – volumi ; 38 cm. ((Settimanale. – Il complemento del titolo varia: periodico della Consociazione repubblicana delle Marche (1914); periodico repubblicano fondato nel 1870 (1964). – Diretto fino al 1904 da Domenico Barilari. - La tipografia varia: Stabilimento tip. cooperativo (1914); Tip. Bellomo (1964). - Descrizione basata su: anno 2, n. 27 (agosto 1871). – Il formato varia: 50 cm (1964). - Copia digitale anni 1914-1918 a: http://www.14-18.it/periodici/AFM_OM_B60_FASC184. – Da anno 146, n. 1 (ott.-dic. 2016) disponibile anche in Internet a: <https://www.luciferonline.it/>. - TO00188040; URB0934447; IEI0163814. Dal 2016 has title: *Lucifero nuovo

***Radio orario** : periodico settimanale / organo ufficiale della Unione radiofonica italiana. - Anno 1, n 1 (18 gennaio 1925)-anno 2, n. 4 (23 gennaio 1926). - Roma : La poligrafica nazionale, 1925-1926. - 1 volume : ill. ; 30 cm. ((L. 1.50 il numero. - BNI 1926-904. - CUB0705457

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***Radiorario** : organo ufficiale della U.R.I., Unione radiofonica italiana : tutti i programmi italiani ed esteri della settimana. - Anno 2, n. 5 (30 gennaio 1926)-anno 5, n. 52 (22 dicembre 1929). - Milano : EIAR, 1926-1929. - 4 volumi : ill. ; 30 cm. ((Settimanale. - Il complemento del titolo cambia. - UM10014518

Autore: Ente italiano audizioni radiofoniche

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***Radiocorriere** : settimanale dell'EIAR. - Anno 6, n. 1 (5/11 gennaio 1930)-anno 19, n. 37 (12-18 settembre 1943). - Torino : EIAR, 1930-1943. - 14 volumi : ill. ; 42 cm. ((Il complemento del titolo varia. - Il formato varia. - TO00202876

Autore: Ente italiano audizioni radiofoniche

Soggetto: Radiotrasmissioni - Periodici

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***Segnale radio** : settimanale dell'Eiar / Ente italiano audizioni radiofoniche. - Anno 1, n. 1 (27 ago.-2 set. 1944)-anno 2, n. 17 (22-28 aprile 1945). - Torino : S.I.P.R.A., 1944-1945 (Torino : Tipografia della S.E.T.). - 2 volumi : ill. ((Direttore Cesare Rivelli. - TO00195117

Autore: Ente italiano audizioni radiofoniche

Soggetto: Radiodiffusione - Italia - Periodici

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***Segnale radio** : musica e propaganda radiofonica nell'Italia nazifascista, 1943-1945 / Gioachino Lanotte. - Perugia : Morlacchi editore U. P., 2014. - 387 p. ; 22 cm. - BNI 2015-2626. - LIA0965392

Fa parte della collezione: *Storia

Autore: Lanotte, Gioachino

Soggetto: Fascismo - Propaganda radiofonica - Ruolo [della] Musica - Italia - 1943-1945

Classe: D384.540945

***Radiocorriere / Radio audizioni Italia. - Ed. per l'Italia centro-meridionale.** - Anno 1, n. 1 (novembre 1945)-anno 3 (1947). - Roma : Rai, Radio Audizioni Italia, 1945-1947. - 3 volumi in folio. ((Settimanale. - BNI 1949-2985. - CFI0362950

***Radiocorriere** : organo ufficiale della radio italiana. - Anno 23, n. 1 (6-12 gennaio 1946)-anno 35, n. 18 (4-10 maggio 1958). - Torino : S.I.P.R.A., 1946-1958 (Torino : S.E.T.). - 13 volumi : ill. ; 42 cm. ((Il complemento del titolo varia. - Il formato varia. - TO00202876

Variante del titolo: *Radio corriere

Soggetto: Radiotrasmissioni - Periodici

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***Radiocorriere TV.** - Anno 35, n. 19 (11/17 maggio 1958)-anno 62, n. 49 (dicembre 1985). - Torino [etc.] : [Edizioni radio italiana], 1958-1985. - 28 volumi : ill. ; 35 cm. ((Settimanale. - Il formato varia. - BNI 58-9386. - RAV0024443

Copia digitale a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

***TV radiocorriere.** - Anno 62, n. 50 (dicembre 1985)-anno 72, n. 53 (31 dicembre 1995); anno 69 (1999)- . - Roma : Nuova Eri, 1985-2008. - volumi : ill. ; 28 cm. ((Settimanale. - Direttore Willy Molco. - CFI0398854

Ha come supplemento: *Italiana [PE. 11647]

Copia digitale 1985-1995 a: <http://www.radiocorriere.teche.rai.it/Default.aspx>

Digitization makes it possible to integrate the collections owned by the library with the missing issues, freely available online, of other libraries.

A new season of cataloguing must be launched starting from the mass of documents placed on the web in recent years and freely available to users. It will concern the (few) publications not yet described, but above all it, will make significant some descriptions already present in the online catalogs.¹⁹

The availability of digital reproductions, especially of old publications, makes it possible to reconstruct the evolution of the titles of a publication correctly, recording the titles and consistencies of the various series.

To obtain a more accurate bibliographic record, it is sometimes necessary to unify information scattered across multiple descriptions. Some information can be derived directly from digitized copies. The digital copy of the original can give us back a record uncontaminated by the use of later printed reproductions.

The comparison of different editions of digital copies can reveal or confirm the presence of parallel publications, not detected in the historical cataloguing, even of important journals.²⁰

In some cases, it will be necessary to establish connections that were non-existent in the catalogs and to create descriptions with the correct serial nature.²¹

Through the analytic cataloguing and filing of the various digital libraries, it is possible to reconstruct a more complete history of the publications, starting from the observation and comparison of different issues of the same publication that may be in different cases and not communicating

¹⁹ For example, going from a description like this: Il *consigliatore : giornale politico, istruttivo, letterario e commerciale. - Pinerolo, 1849-1850. - TO00182029 to a description like this: Il *consigliatore : giornale politico, istruttivo, letterario e commerciale. - Anno 1 (1849)-anno 2, n. 18 (22 febbraio 1850). - Pinerolo : Tipografia Lobetti-Bodoni, 1849-1850. - 18 volumi. ((Settimanale. - Poi: giornale della città e provincia di Pinerolo. - Direttore: Lorenzo Giribaldi. - Descrizione basata su: Anno 1, n. 3 (10 novembre 1849). - TO00182029. Copia digitale a: <https://www.giornalidelpiemonte.it/edizionitesta.php?testata=Consigliatore>

²⁰ La *voce. - Edizione politica. - Anno 7, n. 1 (7 maggio 1915)-anno 7, n. 14 (dicembre 1915). - Roma : Libreria della Voce, 1915. - 14 volumi ; 26 cm. ((Bimensile. - Direttori: Giuseppe Prezzolini; poi: A. De Viti De Marco. - Copertina di colore giallo. - Copia digitale a: <https://fondazionefeltrinelli.it/fonte/la-voce-edizione-politica-1915/#top>. - TO00197733 Variante del titolo: La *voce. Edizione politica. Autore: Prezzolini, Giuseppe

²¹ From the digital library of INEA, the National Institute of Agricultural Economics, we have developed this example: *L*annata agricola ... nel Veneto : prime valutazioni* / Andrea Povellato. - 1988-2000. - Padova : Osservatorio di economia agraria per il Veneto ed il Trentino Alto Adige, 1989-2001. - 13 volumi ; 24 cm. ((Annuale. - Poi: INEA, Istituto nazionale di economia agraria, Osservatorio di economia agraria per il Veneto. - I curatori variano. - CFI0521760. Fa parte di: *Pubblicazioni a cura dell'Osservatorio di Economia Agraria per il Veneto. Autore: Bortolozzo, Davide; Cesaro, Luca; Gambarin, Luigi; INEA; Kuehl, Gerhard; Osservatorio di politica agraria per il Veneto; Povellato, Andrea; Schiavon, Stefano <1971->. Copia digitale: -1994-1998, 2000 a: http://dSPACE.crea.gov.it/handle/inea/1032/browse?type=dateissued&submit_browse=Data+di+pubblicazione -1999 a: <http://dSPACE.crea.gov.it/bitstream/inea/1269/1/VEN-19.pdf>
*L*andamento del settore agroalimentare nel Veneto : prime valutazioni per il ...* / [Andrea Povellato, Stefano Schiavon, Mauro Capriotti, Filippo Codato]. - 2001-2002. - Legnaro (Pd) : Veneto Agricoltura, 2002-2003. - 2 volumi : ill. ; 24 cm. ((Annuale. - Sul frontespizio: Veneto Agricoltura, in collaborazione con Inea. - PUV0880096; PUV0946606. Autore: Povellato, Andrea. Disponibile anche in Internet a: http://dSPACE.crea.gov.it/handle/inea/1235/browse?type=dateissued&submit_browse=Data+di+pubblicazione. *Prime valutazioni ... sull'andamento del settore agroalimentare Veneto / Veneto Agricoltura ; in collaborazione con INEA, Osservatorio economico per il sistema agroalimentare e lo sviluppo rurale. - 2003-2008. - Legnaro PD : Veneto Agricoltura, 2004-2009. - 6 volumi ; 24 cm. ((Annuale. Autore: INEA; Veneto agricoltura. Disponibile anche in Internet a: http://dSPACE.crea.gov.it/handle/inea/904/browse?type=dateissued&submit_browse=Data+di+pubblicazione

with each other.²² We may digitize supplements, without reference to the mother journal, of which we don't even know the bibliographical description.

By comparing digitizations with catalogs and other repertories on the periodical press, we can better define the number of digitized issues compared to those published.²³ The matching between images and bibliographic descriptions must be precise, otherwise we risk keeping publications that are usable hidden. Error is always around the corner, especially in the case of publications with the same title and from the same period.

Through digital copies, we can correct erroneous information in the catalog, related to numbering and possible relationships with homogeneous periodicals. The bibliographical investigation allows us to uncover anomalies in periodicals issues and particular numbering systems.

The comparison of digitizations and previous bibliographical descriptions allows us to determine the periodicity of publications. From the reading of editorials, we can also detect cessations of periodicals and changes of titles.

Assessment elements for a quality digital library

The survey of the 73 digital libraries visited this year has allowed us to define what should be the quality criteria for a national digital library. A ranking was compiled that identifies fourteen criteria:

1. display
2. graphics
3. quality of the alphabetical sorting by titles
4. simplicity, speed, and effectiveness of the search
5. presence (or not) of a bibliographic description of the digitized material
6. presence (or not) of a bibliographic history of the publication
7. linking between the various titles of the publication
8. information about digitized holdings
9. accuracy and precision of the information
10. information about the number of digitized volumes
11. quality of the image display system
12. quality of the images
13. rarity and value of the collections
14. completeness of the digitized collections

²² To reconstruct the history of *The Worker of Trieste* we consulted: the SBN OPAC, Internet culturale website, Biblioteca Attilio Hortis of Trieste website, *Stampa clandestina* website, Wikipedia and Archivio della Federazione di Trieste del Partito della Rifondazione comunista:

<http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8331>

<http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8332>

<http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8335>

<http://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/8336>

http://www.stampaclandestina.it/?page_id=116&ricerca=253

<http://www.rifondazionecomunistatrieste.org/archivio.htm>

²³ *La guerra : pubblicazione settimanale, illustrata*. - Anno 1, n. 1 (27 giugno 1915)-n. 13 (1915). - Roma : Quattrini, 1915. - 1 volume : ill. ; 36 cm. ((BNI 1915-7778. - Copia digitale dei n. 1-10 a: <http://www.14-18.it/periodici/CFI0355788/1915>. - CFI0355788. Soggetto: Guerra mondiale 1914-1918.

The Digital Newspaper Library of BNCR

The Digital Newspaper Library of the National Central Library of Rome (BNCR) will ideally host the bibliographic record and be identified by an alphanumeric code.²⁴

Since it participated in the first European digitization projects, the BNCR has started a constant process of digitization of its collections, increased with materials coming from the participation in European projects and the collaboration with other libraries, organizations, Italian and international institutions. Among the main ones, we recall the Europeana 14-18 Project,²⁵ which has provided for the digitization of 20,000 images of periodicals and historical newspapers; the GoogleBooks Project which,²⁶ under the coordination of BNCR for Italy, has led to the digitization of over 60,000 volumes of periodicals from the period between 1668 and 1946, merged in the collections of the Digital Newspaper Library; a five-year agreement, signed in 2017 between the National Library and the Library of the Senate of the Republic “Giovanni Spadolini”,²⁷ on the implementation of the National Newspaper Library as a single portal of access to the digitized collections of historical newspapers and journals belonging to the two libraries.

With its 2,230 titles of newspapers, periodicals, and historical journals and a patrimony of over 18 million images, the BNCR Digital Newspaper Library represents one of the richest digital newspaper libraries available on the Italian scene, continuing a long historical tradition that has involved the Biblioteca Nazionale centrale di Roma since 1908 with the task of establishing and preserving the National Newspaper Library.²⁸

The available titles are based on METS for the encoding of all descriptive, administrative and structural metadata for the management of digital objects. The creation of an intermediate level, between the list of titles and the list of available years, containing a tab for each record would allow the management of bibliographic and technical information according to a Dublin Core schema.

²⁴ <http://digitale.bnc.roma.sbn.it/tecadigitale/emeroteca/classic>

²⁵ <http://www.14-18.it/>

²⁶ <http://www.bnrcrm.benculturali.it/it/832/progetto-googlebooks>.

²⁷ <http://digitale.bnc.roma.sbn.it/tecadigitale/progettoConvenzioneBS>

²⁸ Andrea De Pasquale, *Per un'emeroteca nazionale digitale*, «Bibliothecae.it», 7 (2018), n. 2: 348-370, <<https://bibliothecae.unibo.it/article/view/8951>>.

Proposed structure for a national digital newspaper library

The structure that we imagine has as a basis a super-record of the Work²⁹ marked by a unique and univocal alphanumeric code, on the model of Wikipedia. It is necessary to avoid the proliferation of descriptions for the same publication.

SEARCH MASKS (Access to Work)

FIRST SEARCH MASK

1. Search by Title. Browse a list of titles
2. Search by author. Browse a list of authors (Authority file)
3. Search by subject. Browse a list of subjects (Thesaurus)

The lists are sorted alphabetically and asyndetically, i.e. by significant word excluding articles and also conjunctions and prepositions if they are not at the beginning of the title. The lists can be divided into 26 blocks corresponding to the letters of the alphabet.

Search by title	Search by author	Search by subject
<i>Antologia</i>	Gabinetto scientifico letterario G. P. Vieusseux	Arte
<i>Nuova antologia</i>	Protonotari, Francesco	Cultura
<i>Nuova antologia di lettere, scienze ed arti</i>	Spadolini, Giovanni	Letteratura
<i>Nuova antologia di Scienze lettere ed arti</i>	Vieusseux, Giovan Pietro	Scienze

Example 1. Search channels

The 12 search channels are all connected to the super-record that we will call **IT2**.

The elements that the super-record should contain are:

- a) Bibliographical description
- b) Digitized volumes with links to the digital libraries
- c) Historical and bibliographical information
- d) Notes and bibliographical references
- e) Technical notes on digitization

²⁹ See IFLA, *Functional requirements for bibliographic records. Final report*, 1998.

A. Bibliographical description

**Antologia*. - Tomo 1, n. 1 (gennaio 1821)-vol. 48, n. 144 (dicembre 1832). - Firenze : al Gabinetto scientifico e letterario di G. P. Vieusseux, 1821-1832. - 48 volumi ; 22 cm. ((Mensile. - Dal 1831 ha il complemento del titolo: giornale di scienze, lettere ed arti. - Disponibile anche in Internet come banca dati e copia digitale a: <http://www.antologia-vieusseux.org/>. - ISSN 1125-3622. - LO10020689

Autore: Gabinetto scientifico letterario G. P. Vieusseux

Soggetti: Arte – Periodici; Letteratura – Periodici; Scienze - Periodici

**Indice generale alfabetico delle materie contenute nell'Antologia, giornale fiorentino diretto da Gio. Pietro Vieusseux* : 1821-1832. - Firenze : A. Cecchi, 1863. - 270 p. ; 23 cm. - CFI0557156

**Nuova antologia di scienze, lettere ed arti*. - Vol. 1, fasc 1 (31 gennaio 1866)-vol. 30, fasc. 12 (dicembre 1875); 2. serie, vol. 1, fasc. 1 (gennaio 1876)-vol. 54, fasc. 24 (16 dicembre 1885); 3. serie, vol. 55, fasc. 1 (1 gennaio 1886)-vol. 60, fasc. 24 (15 dicembre 1895); 4. serie, vol. 61, fasc. 1 (1 gennaio 1896)-vol. 84, fasc. 672 (16 dicembre 1899). - Firenze : Direzione della Nuova antologia, 1866-1899. - 84 volumi : ill. ; 24 cm. ((Mensile; bimensile (1878-1880). - Fondata da Francesco Protonotari. - Dal 1876 fasc. hanno doppia numerazione. - L'editore varia. - Indici 1866-1895. - ISSN 1125-3630. - LO10020526

**Nuova antologia di scienze, lettere ed arti : indice generale dei 30 volumi della prima serie : anni 1866-1875*. - Firenze : Direzione della Nuova antologia, 1876. - IV, 128 p. ; 24 cm. - TSA0336581

**Nuova antologia di lettere, scienze ed arti*. - 4. ser., vol. 85, fasc. 673 (1 gen. 1900)-vol. 120, fasc. 816 (16 dic. 1905); 5. ser., vol. 121, fasc. 817 (1 gen. 1906)-vol. 180, fasc. 1054 (16 dic. 1915); 6. ser., vol. 181, fasc. 1055 (1 gen. 1916)-vol. 244, fasc. 1290 (16 dic. 1925); 7. ser., vol. 245, fasc. 1291 (1 gen. 1926)-vol. 246, fasc. 1298 (21 apr. 1926). - Roma : Nuova antologia, 1900-1926. - 160 volumi : ill. ; 26 cm. ((Quindicinale. - Doppia numerazione dei volumi. - Numeraz. dei fasc. progressiva negli anni. - Il vol. 234 errato nella doppia numerazione. - ISSN 1125-3649. - RAV0105511

**Nuova antologia : rivista di lettere, scienze ed arti*. - 7. serie, anno 61, vol. 247, fasc. 1299 (1 maggio 1926)- . - Roma : Nuova antologia, 1926- . - volumi ; 24 cm. ((Quindicinale; la periodicità varia. - Dal fasc. 2125/2126 (gen.-giu. 1978) il sottotitolo varia in: rivista trimestrale di lettere, scienze ed arti / diretta da Giovanni Spadolini. - Il luogo e l'editore variano in: Firenze : Le Monnier. - Indici: 1866-1985. - Copia digitale 1926-1940 a: <http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/RAV0027419>. -RAV0027419

Soggetti: Cultura - Periodici

Classe: D055.1

**Indici per autori e per materie della Nuova antologia* : dal 1931 al 1950 / compilati da Laura Giuliani. - RMS0049318

**Indici per autori e per materie della Nuova antologia* : dal 1866 al 1930 / a cura di Lodovico Barbi. - Rist. anast. - XXIII, 721 p. ; 24 cm.

**Indici 1866-2003* Disponibili in Internet all'indirizzo: <https://nuovaantologia.it/storia-nuova-antologia/testi-in-pdf/>

B. Digitized volumes with links to the digital libraries

**Antologia 1821-1832*: <http://www.antologia-vieusseux.org/>

**Antologia 1821-1832*: http://www.internetculturale.it/it/16/search?q=&searchType=avanzato&channel_creator=%22Gabinetto+scientifico+letterario+G.+P.+Vieusseux%22&channel_contributor=%22Gabinetto+scientifico+letterario+G.+P.+Vieusseux%22&opCha_contributor=OR&opCha_creator=OR&meta_typeLivello=periodico&pag=1

**Antologia 1821-1822; 1826-1832*: <http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/LO10020689>

**Nuova antologia 1926-1940*: <http://digitale.bnc.roma.sbn.it/tecadigitale/giornali/RAV0027419>

**Indici 1866-2003*: <https://nuovaantologia.it/storia-nuova-antologia/testi-in-pdf/>

C. Historical and bibliographical information

Antologia fu una rivista con periodicità mensile, pubblicata a Firenze dal 1821 al 1833, promossa da Giovan Pietro Vieusseux e da Gino Capponi, cui collaborarono molti intellettuali del tempo.

L'indirizzo della rivista fu sempre nazionale, intendendo abbracciare i problemi generali della cultura italiana del periodo. Prima di dar vita alla rivista, Vieusseux aveva istituito, con sede a palazzo Buondelmonti, un "gabinetto scientifico-letterario" (il celebre Gabinetto Vieusseux) che, oltre a far conoscere la stampa italiana e straniera, diventò un luogo di incontri e discussioni. Furono collaboratori dell'*Antologia* quasi tutti gli intellettuali attivi fra il 1821 e il 1831, tra i quali Giuseppe Poerio, Gabriele Pepe, Pietro Colletta, Pietro Giordani, Niccolò Tommaseo, Giuseppe Montanelli, Francesco Domenico Guerrazzi, Carlo Cattaneo e Giuseppe Montani. Vieusseux fu il primo editore che compensò i propri collaboratori. Fino ad allora infatti, in Italia le collaborazioni non venivano retribuite.

Pur accogliendo le istanze più disparate, la rivista vantava un orientamento comune: una preoccupazione pedagogica, che si sviluppava in chiave antirivoluzionaria; una filosofia eclettica, che escludeva però le ideologie radicali dell'Illuminismo; un'idea di "letteratura impegnata" per fini utili. Sulla rivista le questioni letterarie ebbero un posto marginale, mentre ci si occupò sistematicamente di argomenti sociali (storia, diritto, ecc.) ed economici (economia, statistica, ecc.).

Sul numero di novembre-dicembre 1832 due articoli incontrarono i rigori della censura preventiva, uno dei quali conteneva critiche all'Austria. L'uscita fu ritardata al gennaio 1833. Le autorità chiesero al direttore di rivelare i nomi degli autori dei due pezzi. Al rifiuto del direttore di uniformarsi alla decisione governativa, la rivista fu chiusa d'autorità da parte del granduca Leopoldo II di Toscana, su pressione dell'Austria.

L'*Antologia* fu per una decina di anni un elemento centrale della cultura italiana, superando di gran lunga, coi suoi oltre 500 abbonati, il numero di lettori delle riviste milanesi (si pensi al *Conciliatore*): la diffusione delle idee della rivista promosse la nascita di una borghesia liberale in Toscana e contribuì alla formazione del concetto di egemonia culturale

D. Notes and bibliographical references

*Paolo Prunas, *L'«Antologia» di Gian Pietro Vieusseux. Storia di una rivista italiana*, Roma, Società editrice Dante Alighieri, 1906

**Antologia della «Antologia» (1821-1832). Rassegna di una rivista*, a cura di Emiliano Zazo, 2 voll., Milano, Bompiani, 1945

*Umberto Carpi, *Letteratura e società nella Toscana del Risorgimento. Gli intellettuali dell'«Antologia»*, Bari, De Donato, 1974

*Angiola Ferraris, *Letteratura e impegno civile nell'«Antologia»*, Padova, Liviana, 1978

E. Technical notes on digitization

La digitalizzazione della Biblioteca nazionale centrale di Roma è tratta da microfilm.

La digitalizzazione del Gabinetto Vieusseux è iniziata nel 2015.

IT2 super-record

Title	Publication place	Publication date	Author	Subject	Bibliographic record code
<i>Antologia</i>	Firenze	1821-1830	Gabinetto scientifico letterario G. P. Vieusseux	Arte	IT2
<i>Nuova antologia</i>	Roma	1926-	Protonotari, Francesco	Cultura	IT2
<i>Nuova antologia di lettere, scienze ed arti</i>	Roma	1900-1926	Spadolini, Giovanni	Letteratura	IT2
<i>Nuova antologia di scienze lettere ed arti</i>	Firenze	1866-1899	Vieusseux, Giovan Pietro	Scienze	IT2

Second search mask (Database)

The search for individual issues in the digital libraries, especially for newspapers and weeklies that may include thousands of units, should not proceed by overall chronological browsing, but broken down into years, months and days, possibly using predefined chronological grids that make it easier to locate the issues sought.

Example 1: <https://www.giornalidelpiemonte.it/edizionitesta.php?testata=Il%20Biellese>

Il biellese, a biweekly with 1145 pages of a search for individual issues. The chronological browsing is annoying, also because for each search the system brings back to the initial page, and therefore to search for a month of the magazine, it is necessary to search about ten times and each time to browse all the pages of the site.

Example 2: <https://avanti.senato.it/avanti/>

Avanti! from the Senate Library. With just a few steps you get directly to the day you are looking for. It is possible to browse through the header, visualize the list of the digitized years, select the desired year, choose the edition and the month, visualize the first pages of each day of the month with the date highlighted, once a search is carried out it returns to the previous screen.

Example 2. Issue search within the digital libraries

Closing remarks

In the current scenario, after more than 20 years, it is essential to rethink the cultural policies in the digital field, pooling projects, ideas, financial and human resources, overcoming inappropriate attitudes of personal or institutional pride to start building a common path for the development and use of the Italian digital heritage. We need to rediscover the united effort among public and private bodies and institutions that characterized the success of SBN in the 1980s. For the knowledge, diffusion, and valorization of the Italian digital heritage, we need a tool that resembles what SBN represents today for the bibliographic heritage of Italian libraries.

References

De Pasquale, Andrea. 2018. "Per un'emeroteca nazionale digitale." *Bibliothecae.it*, 7 (2018), n. 2: 348-370. DOI 10.6092/issn.2283-9364/8951

Formenti, Carlo. *Corriere della sera*, 20 novembre 2000, 27.

Guerrini, Mauro. 1999. "Catalogare le risorse elettroniche: lo standard ISBD(ER)." *Biblioteche oggi*, 17 (1999), n. 1: 62.

IFLA. 1998. *Functional requirements for bibliographic records. Final report*. Munchen: K.G. Saur. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

Mangiaracina, Silvana, and Morrioni Cristina. 2018. *Quanto costa l'accesso alle pubblicazioni scientifiche nell'era dell'Open Access?: una prima analisi delle pubblicazioni nel CNR*. Bologna: Biblioteca Area della ricerca di Bologna CNR. <https://zenodo.org/record/1247497#.XoC-JKPOPkU>